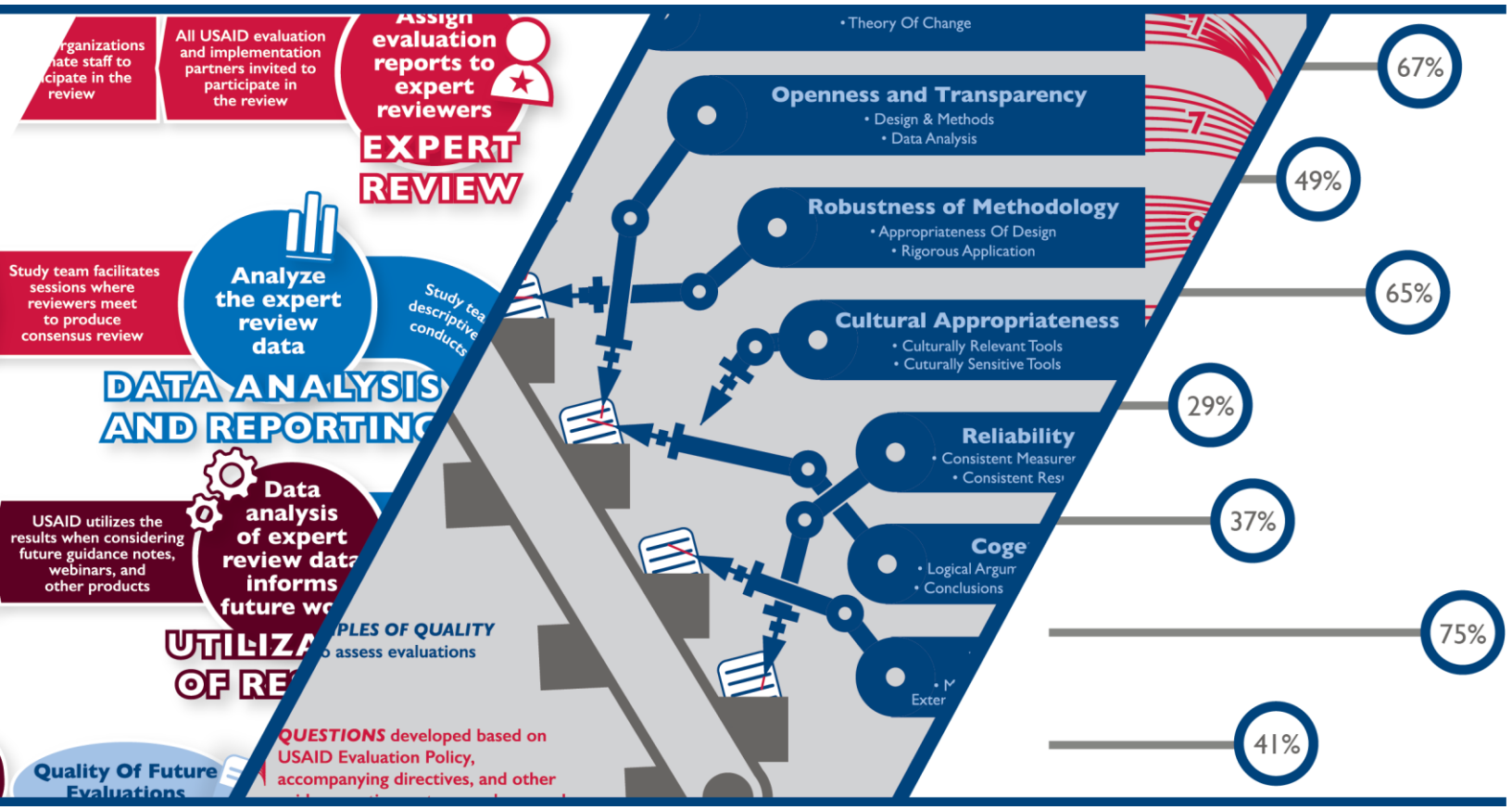
USAID
FROM THE AMERICAN PEOPLE



ASSESSMENT OF THE QUALITY OF USAID-FUNDED EVALUATIONS

Education Sector, 2013-2016

January 2, 2018

This publication was produced for review by the United States Agency for International Development. It was prepared for the E3 Analytics and Evaluation Project and for the Reading and Access Evaluation Project by Management Systems International, A Tetra Tech Company.

ASSESSMENT OF THE QUALITY OF USAID-FUNDED EVALUATIONS

EDUCATION SECTOR, 2013-2016

Contracted under AID-OAA-M-13-00017

E3 Analytics and Evaluation Project

and

under AID-OAA-M-13-00010

Reading and Access Evaluation

Prepared by:

Thomaz Alvares de Azevedo, Team Leader, MSI

ACKNOWLEDGMENTS

This report has been a large collaborative effort to which many individuals contributed their time and expertise. We would like to thank Bhavani Pathak for her leadership as the USAID Contracting Officer's Representative of the E3 Analytics and Evaluation Project and Benjamin Sylla for his leadership as the USAID Contracting Officer's Representative of the Reading and Access Evaluation.

We reserve a special thanks to the expert reviewers for coming together as a community and generously contributing their valuable time. This study would not have been possible without the contributions of: Ellen Bobronnikov and Melissa Chiappetta (Abt Associates), Moses Ngware (African Population and Health Research Center), Elizabeth Spier (American Institutes for Research), Jordene Hale and Laura Harrington (Chemonics), Karen Tietjen (Creative Associates), Alice Michelazzi and Mary Faith Mount-Cors (EdIntersect), Brittany Hebert (Education Development Center), Ania Chaluda and Yvonne Cao (FHI 360), Sean Kelly (Georgetown University), Ed Allan (International Business and Technical Consultants, Inc.), Christine Allison (JBS International), Stefany Thangavelu (Juárez & Associates), Emilie Bagby and Nancy Murray (Mathematica Policy Research), Gaelle Simon, Jennifer Shin, Jeff Davis, Mai Yang, Marie-Louise Orsini, Nitika Tolani, Rouba Reaidi (Management Systems International), Christine Beggs (Room to Read), Amber Gove, Amy Mulcahy-Dunn, Emily Kochetkova, Jennae Bulat, Jonathan Stern and Tracy Brunette (RTI International), Aimee Reeves, Fernanda Gandara, Hetal Thukral, Jasmina Josic and Louise Bahry (School-to-School International), Deepa Srikantaiah (University Research Co.), Elena Stacy (USAID), Haiyan Hua (World Education), Meri Ghorkhmazyan (World Learning), Maria Brindlmayer (YouthPower Learning), Clare Ignatowski, Jeremy Kock, Kristen Potter, Michael Costello, Rosemary Taing and Shamima Tasmin (independent consultants). We also thank Luis Crouch and Varuni Dayaratna who, while unavailable to participate in the review process, still found the time to share their thoughts and expertise. We also thank Clare Ignatowski, Jeff Davis, and Nitika Tolani, who will lead the synthesizing of findings and lessons learned during Phase 2 of this study. Elena Stacy, Heather Risley, and Lauren Greubel from USAID graciously moderated the discussion during the full-day reviewer meeting at MSI. Finally, we thank Elena Walls from USAID for her support and guidance throughout the entire process.

Thomaz Alvares de Azevedo of MSI oversaw this study. In addition, the study greatly benefited from the support of MSI project managers Jeremy Gans and Haley Barry, software developer Chase Gruber, and designer Adam Bloom. We also thank Amanda Kitanga, Elena Szost, Laura Sinclair, Meredith Waters, and Zuhayr Ahmed for their support on many critical needs.

CONTENTS

List of Figures

ACRONYMS

ADS	Automated Directives System
BE^2	Building Evidence in Education
CEA	Cost-Effectiveness Analysis
CIES	Comparative and International Education Society
DEC	Development Experience Clearinghouse
E3	Bureau for Economic Growth, Education, and Environment (USAID)
EiCC	Education in Conflict and Crisis
FY	Fiscal Year
GAO	Government Accountability Office
IRB	Institutional Review Board
MSI	Management Systems International
OCA	Organizational Capacity Assessment
RERA	Rapid Education and Risk Analysis
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Study Background, Objectives, and Limitations

The Office of Education in the United States Agency for International Development's Bureau for Economic Growth, Education, and Environment (USAID/E3) commissioned a team led by Management Systems International to assess the quality of USAID-funded evaluations in the education sector. This request was motivated by the need of the Office of Education to curate, analyze, and disseminate the robust evidence generated by USAID related to the objectives laid out in the Agency's 2011 Education Strategy.

The Office of Education's main objective for this study was to identify areas of strength and weakness in USAID-funded evaluations in the education sector. The study defined "evaluations" in accordance with USAID Evaluation Policy, and reviewed impact and performance evaluations but excluded assessments and informal project reviews.[1] The evaluations included in this study spanned the Agency's three Education Strategy Goals[2] and all six USAID regions,[3] and included evaluations conducted in countries ranging from low- to upper-middle income, and those that were or were not in crisis and conflict. A by-product of this study is a tool to appraise the quality of evaluation reports that is responsive to USAID's cross-sector guidance on evaluations as well as applicable to sector-specific education evaluations. The Office of Education may use findings from this study to identify specific topics on which it could develop additional guidance, products, and presentations to improve the quality of evidence generated for USAID-funded activities in the education sector.

Only evaluations published between 2013 and 2016 that the Office of Education deemed relevant to the Agency's Education Strategy were included in this study. While USAID launched its Education Strategy in 2011, its Implementation Guidance took over a year to finalize and country missions then needed time to align their programs with it. This implies that evaluations reviewed in this study included projects and activities that were undergoing a transitional period, which likely had implications on the design, implementation, and overall quality of evaluations. The study only reviewed evaluations published after the expected date for missions to have programs aligned with the Education Strategy. In addition, there were activities supporting the Education Strategy Goals awarded in 2014-2015, meaning that there will continue to be evaluations relevant to the Education Strategy published through at least 2019.

Study Methods

The Office of Education set the following inclusion criteria for evaluations to be reviewed in this study:

1. USAID-funded evaluations of education interventions;
2. Performance and impact evaluations (additionally, the Office of Education requested the inclusion of a small number of research studies that did not evaluate a specific intervention);
3. Relevant to the Education Strategy;
4. Published between 2013 and 2016;
5. Single, latest published report (in case of reports for multiple phases of an evaluation); and

[1] A substantial number of USAID-funded reports, such as early grade reading assessments that were not tied to an evaluation, were thus beyond the scope of this study.

[2] These evaluations addressed activities related to: (1) improving the reading skills of students in primary grades to increase school success and completion; (2) increasing employment opportunities for youth or strengthening higher education systems so youth can find quality jobs and contribute to the economic growth and peace and stability of their countries; or (3) increasing access to and decreasing dropout from primary and secondary schools, especially in crisis and conflict environments where children and youth depend on improved education service delivery, equity, and safety in schools. The Office of Education determined which evaluations were relevant.

[3] These six regions are: Afghanistan and Pakistan; Africa; Asia; Europe and Eurasia; Latin America and the Caribbean; and Middle East.

6. Evaluation reports from multiple countries (in case of a multi-country education intervention).

In collaboration with the Office of Education, the study team developed an evaluation quality tool based on a framework for assessing principles of quality that was prepared by the United Kingdom's Department for International Development and produced by the Building Evidence in Education (BE^2). The BE^2 framework includes seven principles of quality: (1) conceptual framing, in terms of its theory of change; (2) openness and transparency, in terms of self-criticism and independence; (3) robustness of the methodology, in terms of appropriateness of the design and rigorous implementation; (4) cultural appropriateness, in terms of culturally relevant tools and culturally sensitive analysis; (5) validity, in terms of measurement, internal, external, and ecological validity; (6) reliability, in terms of consistent measurement and results from repeated processing and analysis; and (7) cogency, in terms of logical argumentative thread throughout the entire report and conclusions being based on the evaluation's findings. For each principle, the study team developed assessment items based on the USAID Evaluation Policy[4] and relevant Automated Directives System (ADS) sections for evaluation;[5] the team also adapted items from established evaluation report quality checklists.[6] Prior to this assessment, the study team piloted the tool in a workshop co-presented with the Office of Education at the Comparative and International Education Society's 2017 annual conference.

For the review process, the Office of Education requested that organizations nominate staff to serve as reviewers on this study. This served three purposes: gathering broad feedback on the tool, disseminating the BE^2 framework, and providing an opportunity for community members to read and discuss each other's evaluations. Thirty-six reviewers from 21 organizations participated as volunteer reviewers for this study. The study team developed an online platform for each evaluation to be reviewed by two reviewers. Each pair of reviewers also met virtually to reconcile any differences in scoring and produce consensus responses. The study team provided online training and support, and hosted an event in which reviewers met in person to discuss questions that arose when conducting the reviews as well as to provide feedback on the tool and web platform.

Findings[7]

The study team analyzed data from these reviews and this report discusses the findings. Figure 1 shows the percentage of the 92 evaluations rated as adequately addressing each of the principles of quality. Cogency was most frequently rated as adequate (75 percent of the evaluations) while cultural appropriateness was least frequently rated as adequate (29 percent of the evaluations).

[4] See USAID, *Evaluation Policy* (Washington, D.C.: USAID, October 2016), https://www.usaid.gov/sites/default/files/documents/1870/USAIDEvaluationPolicy.pdf.

[5] ADS 201maa "USAID's Criteria to Ensure the Quality of the Evaluation Report," ADS 201mah "USAID Evaluation Report Requirements," and ADS 201sae "USAID Data Quality Assessment Checklist and Recommended Procedures." Also, USAID Scientific Research Policy.

[6] E3 Sectoral Synthesis of FY2015 Evaluation Findings, the Critical Appraisal Skills Programme Qualitative Checklist, What Works Clearinghouse's Procedures and Standards Handbook, Running Randomized Evaluations: A Practical Guide, and the Early Grade Reading Assessment Toolkit: Second Edition.

[7] In the second phase of this study, the data will be used to determine which evaluations meet the Office of Education's quality standards for inclusion in the synthesis about topics of interest under each Education Strategy Goal.

FIGURE 1: PERCENTAGE OF EVALUATIONS SCORED AS ADEQUATE BY PRINCIPLE OF QUALITY (N=92)

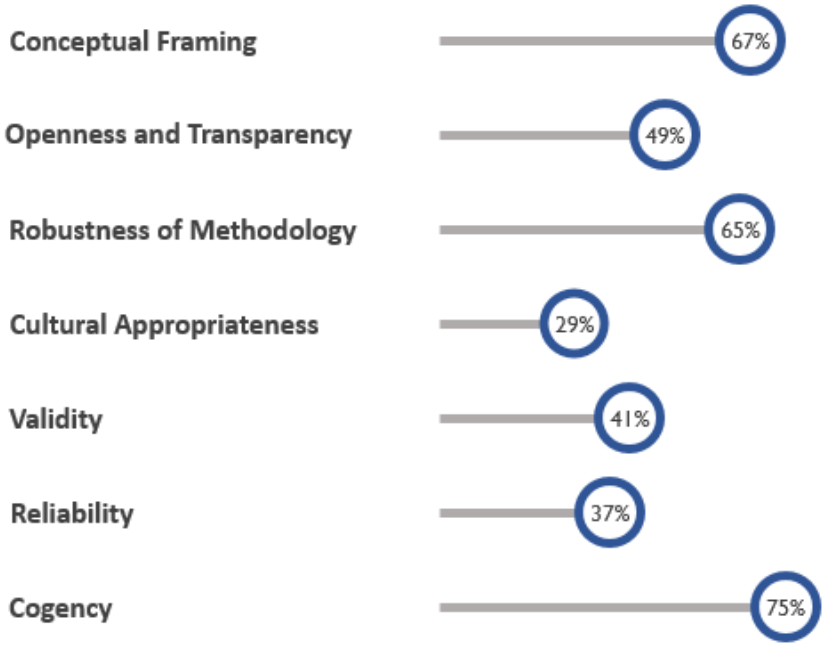
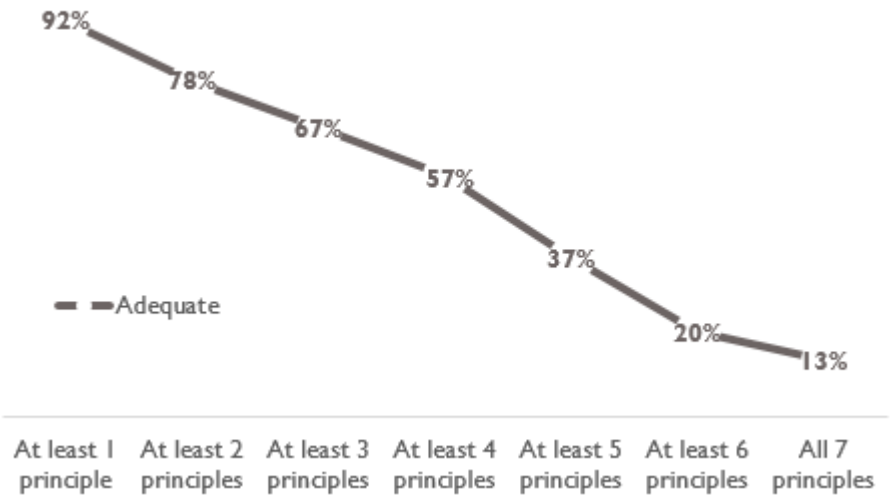


Figure 2 shows the cumulative frequency for the number of principles of quality rated as adequate. About 8 percent of evaluations failed all seven principles of quality, while around 13 of evaluations were adequate on all principles. About 63 percent had at most half of the principles of quality rated as adequate (i.e., four or less).

FIGURE 2: PERCENTAGE OF EVALUATIONS BY NUMBER OF ADEQUATE PRINCIPLES OF QUALITY (N=92)



Discussion and Conclusion

Overall, the evaluations showed greater strength in conceptual framing and cogency, and greater weakness in validity and reliability. This aligns with findings from the U.S. Government Accountability Office's (GAO's) 2017 performance audit about how Agencies Can Improve the Quality and Dissemination of Program Evaluations, which reviewed 49 performance and 14 impact evaluation reports from all USAID sectors. This suggests that the quality issues facing the education sector are present in other sectors. Furthermore, the findings from the GAO report extended beyond USAID, which suggests that this is a struggle for other foreign assistance agencies, which may speak to the difficulties of evaluating programs in challenging environments abroad. While these findings align, this study and the GAO report differed in many technical respects, including the report publication dates, stages of the evaluations, method for double rating, and evaluation quality tool and framework. For example, cultural appropriateness was an area of great weakness for evaluations reviewed in this study, but the GAO report did not assess that principle of quality.

The study team also investigated factors that the Office of Education expected to affect evaluation quality. Neither country income level nor crisis and conflict status was strongly associated with the consensus responses to the evaluation quality assessment, which implies that these are poor predictors of quality for the USAID-funded evaluations in education. The Education Strategy Goal was also a poor predictor of whether the evaluation adequately addressed principles of quality. However, there were strong associations between items in the evaluation quality tool and the type of evaluation (i.e., whether impact, quantitative, or qualitative performance evaluation). These findings suggest two key conclusions:

- Performance evaluations were more likely than impact evaluations to fail to address validity aspects such as measurement, internal, and external validity. This might suggest that the emphasis that the Agency and other donors have placed on improving the quality of impact evaluations has been successful in improving their validity, but this has not yet transferred to performance evaluations.
- The Office of Education's learning agenda could benefit from further consideration of recommendations for qualitative evaluations, especially for leveraging their complementary exploratory and explanatory power with respect to quantitative evaluations – perhaps through sequential data collection in mixed-methods evaluations.

Finally, a positive consequence of this study was that the crowdsourcing process provided an opportunity for the international education community to come together to discuss quality standards for USAID-funded evaluation with the Office of Education. Reviewers expressed concerns that, if this process becomes more established, the evaluation quality tool might become another burdensome requirement. However, many reviewers supported the process used for this evaluation quality assessment being repeated periodically. Reviewers also mentioned that this process provided an opportunity for experts to read each other's evaluations, which led to constructive discussions about quality standards and the subject matter of the reviewed studies.

This study demonstrates the benefits of assessing the quality of USAID-funded evaluations in the education sector. The holistic framework built on the BE^2 working group proposal, mapping different aspects of an evaluation to seven principles of quality. While the items and item descriptors for the evaluation quality tool may be further revised based on feedback from members of the international education community, the results from this initial review have already provided valuable insights into areas of strength and weakness.

INTRODUCTION

Background

Five years after instituting the Agency's Education Strategy,[8] the Office of Education in the United States Agency for International Development's Bureau for Economic Growth, Education, and Environment (USAID/E3) commissioned a study to explore the quality of evaluations relevant to the Strategy's Goals.[9] This study assessed the quality of 92 evaluation reports.[10]

This study included both performance and impact evaluations, based on the definitions in USAID Evaluation Policy.[11] Performance evaluations assess the extent to which a project or activity operates as intended (i.e., process evaluations) or the extent to which it achieves its outcome-oriented objectives (i.e., outcome evaluations). Impact evaluations assess the net effect of a project or activity by comparing the outcomes of interventions with an estimate of what would have happened in the absence of the interventions. The Office of Education also requested the inclusion of a few research studies that did not evaluate a specific intervention.

Given USAID's directions in the Implementation Guidance to the 2011 USAID Education Strategy,[12] which set the target date for Missions to have programs aligned with the strategy as the beginning of FY13, the Office of Education requested that this study include only evaluation reports published between 2013 and 2016. The Office of Education has highlighted its interest in repeating the assessment of evaluation quality in future years as well.

When an evaluated project or activity had reports for multiple phases (e.g., baseline, midterm, final), the study included only the latest published report. When a project or activity was implemented in several countries, the study included evaluation reports for each of the countries. The study included only evaluations the Office of Education considered relevant for the Education Strategy. All evaluations were funded by USAID.

Members of the international education community generously contributed their valuable time and expertise to this study. Thirty-six representatives from 21 organizations reviewed evaluations, with the remaining evaluations reviewed by Management Systems International (MSI) staff and consultants.

A team led by MSI is conducting this study across two mechanisms: the E3 Analytics and Evaluation Project (implemented by MSI in partnership with Development and Training Services, a Palladium

[8] See: http://pdf.usaid.gov/pdf_docs/PDACQ946.pdf.

[9] Education Strategy Goals: (1) improved reading skills for 100 million children in primary grades, (2) improved ability of tertiary and workforce development programs to generate workforce skills relevant to a country's development goals, and (3) increased equitable access to education in crisis and conflict environments for 15 million learners. Most of USAID's current education programming targets one or more of these Goals, and evaluations of its activities are routinely conducted. See: http://pdf.usaid.gov/pdf_docs/PDACQ946.pdf

[10] The study team identified education evaluations based on previous synthesis reviews conducted by Management Systems International, complemented by searches in the Development Experience Clearinghouse (DEC), and the communities of practice associated with each of the three Education Strategy Goals. The Office of Education then vetted the final list of evaluations.

[11] Evaluation is the systematic collection and analysis of information about the characteristics and outcomes of strategies, projects, and activities as a basis for judgments to improve effectiveness, and timed to inform decisions about current and future programming. See: https://www.usaid.gov/sites/default/files/documents/1870/USAIDEvaluationPolicy.pdf

[12] See: http://pdf.usaid.gov/pdf_docs/Pdact461.pdf

company; and NORC at the University of Chicago) and Reading and Access Evaluation (implemented by NORC with MSI as a subcontractor). Annex 1 provides USAID's statement of work for this study.[13]

Objective and Intended Audience

As stated in the USAID Evaluation Policy, the primary purposes of USAID-funded evaluations are both accountability (e.g., whether the project is working) and learning (e.g., what would it take for the project to be replicated in another time or context). This focus on using the evaluations' learning to inform future programming and ultimately improve development efficiency has been embraced by USAID in guidance documents such as USAID's Collaborating, Learning, and Adapting (CLA) Toolkit.[14]

Because not all evidence is created equal, the Office of Education needed a tool to appraise the quality of USAID-funded education evaluations. This tool should be responsive to USAID's evaluation policies and guidance documents (i.e., cross-sector), as well as tailored towards education (i.e., sector-specific). Once this tool was used to appraise individual evaluations, the Office of Education also needed the collective results of the assessment to be amenable to a framework that broke down quality into several domains, which would allow the Office of Education to identify areas of strength and weakness in the body of evidence and consider needs for future guidance notes, webinars, and other products to improve the quality of evidence generated by USAID in the education sector.

The Office of Education's main objective for this study was to identify areas of strength and weakness in USAID-funded evaluations in the education sector. This entailed developing a tool for appraising the quality of evaluation reports that is responsive to USAID's cross-sector guidance on evaluations and is sector-specific to education evaluations,[15] and ensuring that the information resulting from the application of this tool to a multitude of evaluation reports can be used to identify areas of strength and weakness in the evaluations funded by USAID in the education sector. The study team will also use the results of the assessment of the quality of evaluations to identify the evaluation reports that meet the Office of Education's minimum quality standards, for inclusion in a second phase of the study in which the team will synthesize findings and lessons learned about topics of interest to the Office of Education under each Education Strategy Goal.

The primary audiences for this study are USAID/E3 Office of Education and USAID Mission staff, as well as implementing and country partner organizations that plan and deliver education and workforce development programs and related support services.

METHOD

Selection Criteria

The Office of Education established the following inclusion criteria for this study. Annex 2 describes the process used to identify the evaluation reports.

[13] Annex 1 provides the statement of work for the E3 Analytics and Evaluation Project's component of this study, focusing on Goal 2. The Reading and Access Evaluation project's statement of work for this study is nearly identical, replacing references to Goal 2 with Goals 1 and 3.

[14] Integrating CLA practices into the work helps to ensure that programs are coordinated with others, iteratively adapted to remain relevant throughout implementation, and grounded in a strong evidence base. See: https://usaidlearninglab.org/cla-toolkit

[15] The tool was developed in accordance with internationally accepted frameworks for appraising the quality of education research set by the Building Evidence in Education (BE^2) donor working group, for which USAID is part of the Steering Committee. The holistic framework proposed may be suitable to other sectors as well.

- USAID-funded evaluations of education interventions;
- Performance and impact evaluations (additionally, the Office of Education requested the inclusion of a small number of research studies that did not evaluate a specific intervention);
- Relevant to the Education Strategy;[16]
- Published between 2013 and 2016;
- Single, latest published report (in case of reports for multiple phases of an evaluation); and
- Evaluation reports from multiple countries (in case of a multi-country education intervention).

This study included 92 evaluation reports. The Office of Education vetted the final list of evaluations.

Instrument Development

The Office of Education requested that the evaluation quality tool used in this study meet the following requirements:

- Be in accordance with USAID guidance pertaining to evaluations;
- Be in accordance with internationally accepted frameworks for appraising the quality of education research;
- Not be biased in favor of any particular type of evaluation (impact or performance) or research methods (quantitative or qualitative);
- Be amenable to USAID's heterogeneous set of evaluation questions; and
- Balance the length of the tool (number of items) with the breadth of the framework (number of principles of quality used).

To buttress the type of learning sought by the Office of Education, the tool also sought to capture information about what happened between the intervention and the outcome, such as the theory of change behind the project or activity being evaluated, whether the local conditions held for that theory to apply, how strong the evidence was for the behavior change expected by the project or activity, and what the evidence was that the implementation process was carried out well.[17]

The study team developed items for the tool that were grounded in USAID guidance regarding evaluation reports;[18] the team also adapted items from established evaluation report references and quality checklists.[19] The team then mapped all items to the internationally agreed framework for assessing the quality of education evaluations outlined by the guidance note prepared by the United Kingdom's Department for International Development and produced by BE^2 [20] on Assessing the Strength

[16] These evaluations addressed activities related to: improving the reading skills of students in primary grades to increase school success and completion; increasing employment opportunities for youth or strengthening higher education systems so youth can find quality jobs and contribute to the economic growth of their countries; or increasing equitable access to education in crisis and conflict environments. The Office of Education determined which evaluations were relevant.

[17] Adapted from the questions about generalizability found in Glennerster, Rachel, and Kudzai Takavarasha, *Running Randomized Evaluations: A Practical Guide*. New Jersey: Princeton University Press, 2013. See: http://runningres.com/

[18] This guidance includes USAID Evaluation Policy, USAID Scientific Research Policy, and relevant Automated Directives System (ADS) sections for evaluation including ADS 201maa "USAID's Criteria to Ensure the Quality of the Evaluation Report," ADS 201mah "USAID Evaluation Report Requirements," and ADS 201sae "USAID Data Quality Assessment Checklist and Recommended Procedures."

[19] Other sources include the E3 Sectoral Synthesis of Evaluation Findings, the Critical Appraisal Skills Programme Qualitative Checklist, What Works Clearinghouse's Procedures and Standards Handbook, Running Randomized Evaluations: A Practical Guide, and the Early Grade Reading Assessment Toolkit: Second Edition.

[20] BE^2 is a donor working group started in 2012 that includes 30 member organizations, led by USAID, the World Bank, the United Kingdom's Department for International Development, and multiple United Nations agencies. It promotes evidence to

of Evidence in the Education Sector.[21] The framework consists of seven principles of quality: the conceptual framing of the study, its openness and transparency, the robustness of the methodology, the cultural appropriateness of the tools and analysis, the validity and reliability of the findings, and the cogency of the report. Unlike other evidence rating systems, such as the What Works Clearinghouse,[22] Clearinghouse of Labor Evaluation and Research,[23] and EVIRATER,[24] this study's tool assessed principles of quality for the overall evaluation instead of for individual findings, which is similar to what the U.S. Government Accountability Office (GAO) did in its performance audit on how Agencies Can Improve the Quality and Dissemination of Program Evaluations.[25]

The study team and the Office of Education piloted the evaluation quality tool and then co-presented it at a workshop during the 2017 annual conference of the Comparative and International Education Society (CIES). During this workshop, attendees from USAID implementing and evaluation partner organizations, as well as from universities, re-piloted the tool and provided feedback.

After the CIES conference, the study team worked with the Office of Education to incorporate this feedback into the tool, including shortening it to 40 core questions (4 to 8 questions per principle of quality) plus an overall expert judgment of adequacy and accompanying justification for each principle, resulting in 54 questions. The team tested the tool with a larger set of experts from the international education community during the review process, and the Office of Education will use the feedback from the expert reviewers (summarized in the "Findings" section and addressed in the "Discussion" section of this report) to enhance the evaluation quality tool for future application.

In collaboration with the Office of Education, the study team also produced an 18-question tool for capturing background information from the evaluation reports about activity and evaluation characteristics. The study team and the Office of Education piloted this tool, and only MSI reviewers used it.

Figure 3 summarizes the evaluation quality review process that the study team employed. Annex 4 provides the final tools used for this study, the sources for the items that were adapted, and the passages that inspired the items used in the evaluation quality tool. Annex 7 provides the bibliographical information and links to the source materials.

The team incorporated the tools into a web platform built using an open-source web application, Ruby on Rails. This platform allowed for evaluations to be reviewed online. Given that these reviews needed to be integrated with other management components, the team adapted the web platform developed for the E3 Sectoral Synthesis of 2015 Evaluation Findings, which relied on a collaborative review process between the E3 Analytics and Evaluation Project and USAID/E3.[26] The developed web platform facilitated data capture and extraction from PDF reports (i.e., allowing for two panes to be displayed side by side, with the report on one pane and the fields for the data capture application on the other) and had built-in database security features to ensure that the team could protect the identity of the

inform policy and make programming decisions, and build common standards on how to assess evidence from education evaluations.

[21] See: https://www.usaid.gov/sites/default/files/documents/1865/BE2_Guidance_Note_ASE.pdf.

[22] See: https://ies.ed.gov/ncee/wwc/.

[23] See: https://clear.dol.gov/.

[24] See: http://abtassociates.com/Noteworthy/2015/EVIRATER-Rating-the-Strength-of-Evidence-in-Evalua.aspx.

[25] See: http://www.gao.gov/assets/690/683157.pdf.

[26] For the E3 Sectoral Synthesis of FY2015 Evaluation Findings, a team of 61 specialists from 10 E3 offices and 5 MSI team members extracted lessons learned, project results, areas for improvement, and innovative practices as well as cross-cutting topics related to gender equality and women's empowerment, private sector engagement, and governance, from 92 evaluations.

evaluation reviewer as well as maintain the capability to set user permissions. The application also included monitoring features to allow for tracking progress and automated email messaging.

Review Process

The Office of Education and the study team developed a process for reaching out to implementing and evaluation organizations to nominate staff as reviewers for this study. Annex 3 describes the process.

Thirty-six experts from 21 organizations participated as volunteer reviewers for this study. This approach was motivated by discussions with participants at the workshop co-led by the Office of Education and the study team at the CIES 2017 annual conference.

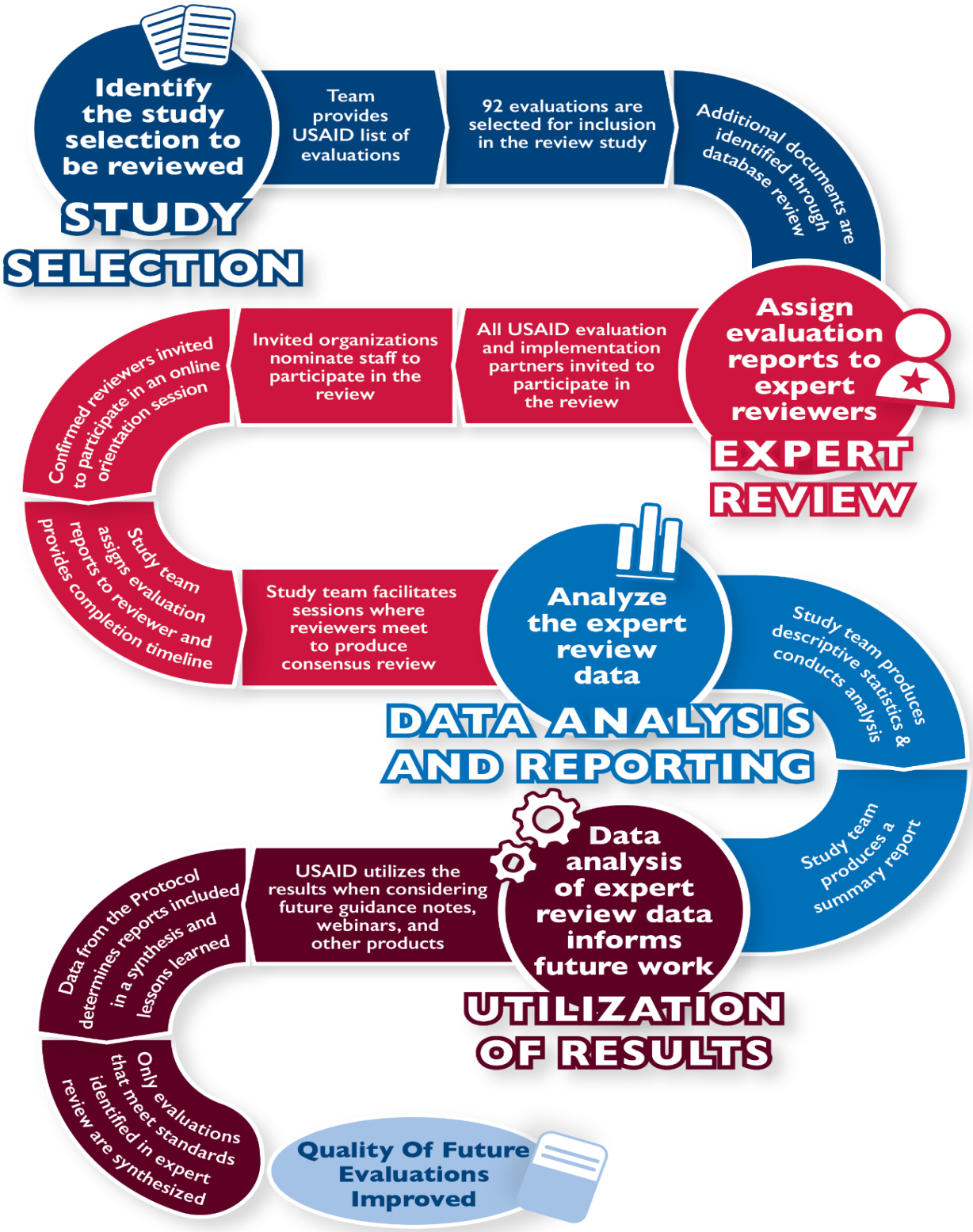
Each expert reviewed two or three evaluations. The study team complemented the volunteer reviewers with eight MSI staff members and six consultants, who also substituted for the volunteer reviewers in case they were unable to complete their assignments.[27] If reviewers could not harmonize their reviews, the study team reviewed the evaluation and served as arbiter. Two USAID staff members also volunteered as reviewers for this study.

The study team took several steps to ensure consistency among reviewers' responses, including providing an orientation package and rater's guide, facilitating reviewers' meetings, and providing remote support. Two experts independently reviewed each evaluation and then met remotely to harmonize their responses.[28]

[27] These include MSI staff and consultants who will work on the second phase of this study to synthesize findings and lessons learned on topics of interest for the Office of Education under each Education Strategy Goal, based on the evaluations that met the Office of Education's quality standards for inclusion. All met the minimum reviewer qualifications for this study.

[28] This parallel review process differs from the sequential review process used in other co-review exercises, such as the GAO's performance audit of evaluation quality, Agencies Can Improve the Quality and Dissemination of Program Evaluations, for which the first reviewer reviewed the evaluation, and the main responsibility of the second reviewer, who had access to the first reviewer's scoring, was to indicate whether he or she agreed with the first reviewer's scoring.

FIGURE 3: SUMMARY OF THE REVIEW PROCESS



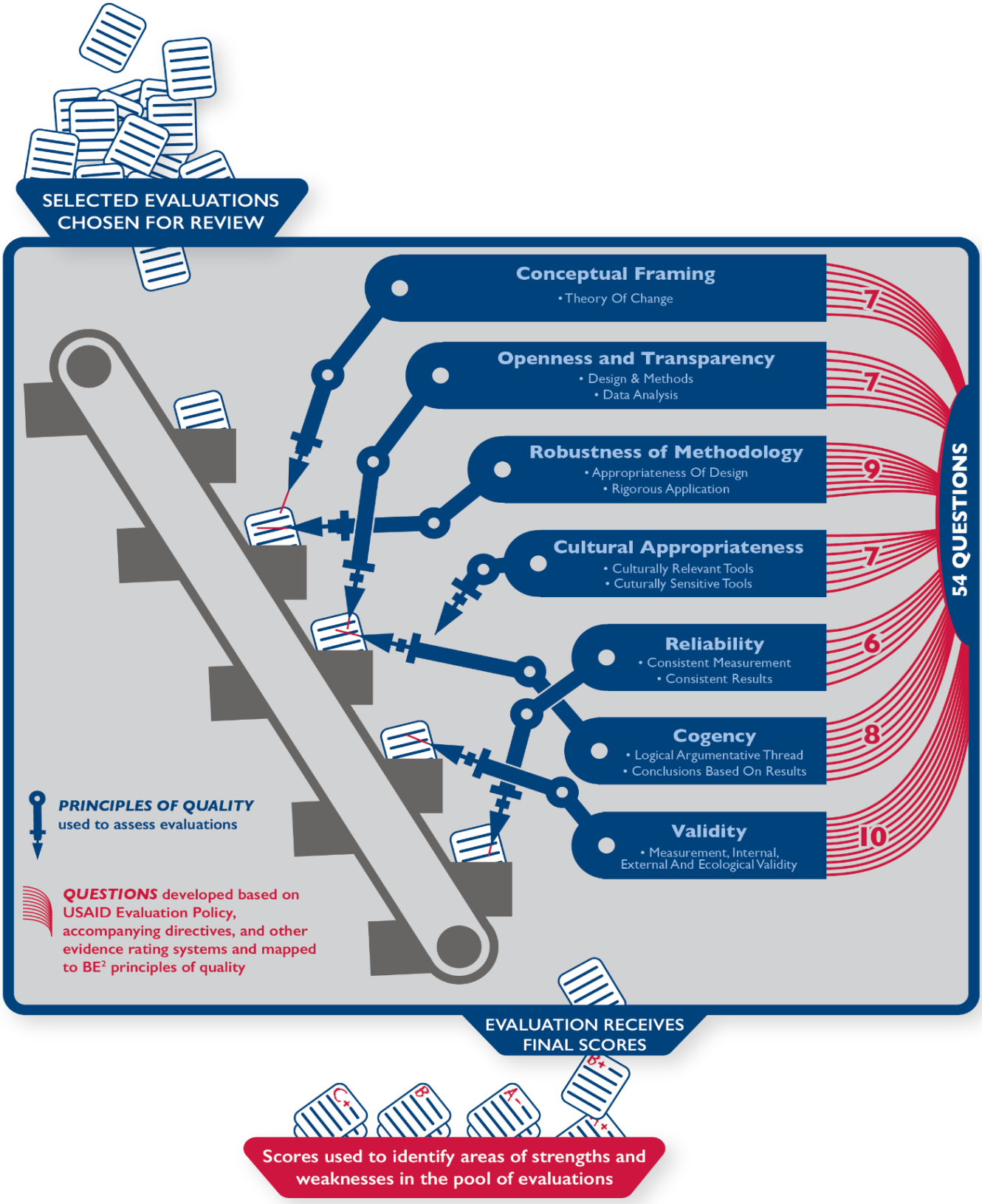
Evaluation Quality Scoring

Reviewers rated "adequate" or "not adequate" to whether the evaluation met standards of quality for each of the seven principles (seven questions) and then provided a brief justification (seven questions). It should be noted that this study did not produce overall quality scores for each evaluation but rather it provides the consensus expert judgement from two reviewers for each principle of quality for each evaluation. As emphasized in the USAID Evaluation Policy, different evaluation design types are appropriate for different types of study questions. Thus, whether the reviewers considered an evaluation "adequate" in terms of each principle of quality was relative with respect to the evaluation type.[29]

Reviewers answered as "yes", "partial" or "no" to 40 questions divided across the 7 principles of quality: conceptual framing (5 questions), openness and transparency (5 questions), robustness of the methodology (7 questions), cultural appropriateness (5 questions), validity (8 questions), reliability (4 questions), and cogency (6 questions). A "partial" score could be given when some but not all elements in an evaluation met a criterion. While most items were applicable to all evaluation types, some were applicable only to specific evaluation types. Annex 4 maps which items were applicable to each evaluation type.

[29] Similar to the GAO's performance audit, this study did not determine a single definition of appropriate or sufficient, because the definition is dependent on the study objectives and data collection conditions.

FIGURE 4: EVALUATION QUALITY SCORING



Data Source and Data Analysis

The study team analyzed the consensus responses using the Stata software package to produce descriptive statistics (i.e., frequency distributions) as well as measure the strength of association for nominal data (Cramer's V). Because all available reports were included in this study, the team did not include chi-square for determining significance. Following discussions with the Office of Education, results for individual evaluation reports are not being made public. However, this report quotes reviewers' judgments of evaluations to provide illustrative examples.

To assess the adequacy of the tool for assessing the quality of the evaluations, the study team and the Office of Education co-hosted a full-day workshop for reviewers. During this workshop, the study team broke reviewers into small groups and led small group discussions about the tool. Reviewers also submitted feedback to the tool online through the web platform developed for this study. The study team conducted a thematic analysis to identify areas of common feedback. Annex 4 provides results from this full-day workshop.

Utilization of Results

The results described in this report provide an initial assessment of the quality of USAID-funded evaluations published between 2013 and 2016 that relate to USAID's Education Strategy Goals. The Office of Education may review these results to explore areas that could benefit from guidance notes, webinars, and other products to improve the quality of the evidence generated by USAID-funded evaluations in the education sector. The Office of Education may share the tool with evaluators of specific projects or activities ahead of the evaluation. The Office of Education has also indicated an interest in periodically repeating the evaluation quality assessment. The study team, in collaboration with the Office of Education, will incorporate the feedback from the expert reviewers for future applications of the quality tool. Finally, the study team will use the results from the present application to determine which evaluation reports meet the standards set by the Office of Education for inclusion in the second phase of this study, which will synthesize findings and lessons learned about topics of interest under each Strategy Goal.

Limitations

The design and implementation of this study faced several limitations, including:

- The design, implementation, and overall quality of the evaluations reviewed in this study were likely affected by the programmatic realignment that the projects and activities being evaluated faced during the rollout of the Education Strategy. Furthermore, this study did not review evaluations of projects and activities supporting Education Strategy Goals that were awarded since 2014, and there will continue to be evaluations relevant to the Education Strategy published through at least 2019. The quality of those evaluations may differ from those reviewed in this study, so this assessment is not necessarily generalizable to all evaluations examining USAID activities that were relevant to Education Strategy Goals.
- Some of the information assessed in the evaluation quality tool might not have been included in the evaluation report since it was not a part of the original evaluation statement of work (e.g., the evaluation questions). Thus, some evaluations may have been assessed negatively for not including information that was not provided to the evaluation team by the commissioning USAID operating unit.
- The Office of Education's guidance for the inclusion of certain information in evaluations related to the education sector, such as documenting the approval of an ethics review for human

subject protection or reporting inter-rater reliability statistics, is new. Evaluations therefore may be assessed negatively for not having included information that was not required or recommended per USAID guidance at that time.

- Reviewers were instructed to apply the evaluation quality tool only to information that was provided in the evaluation reports. The study did not examine any other evaluation or project-related documentation.
- This study did not include value for money among the principles of quality assessed, as examining cost information was outside of the scope of the assessment.

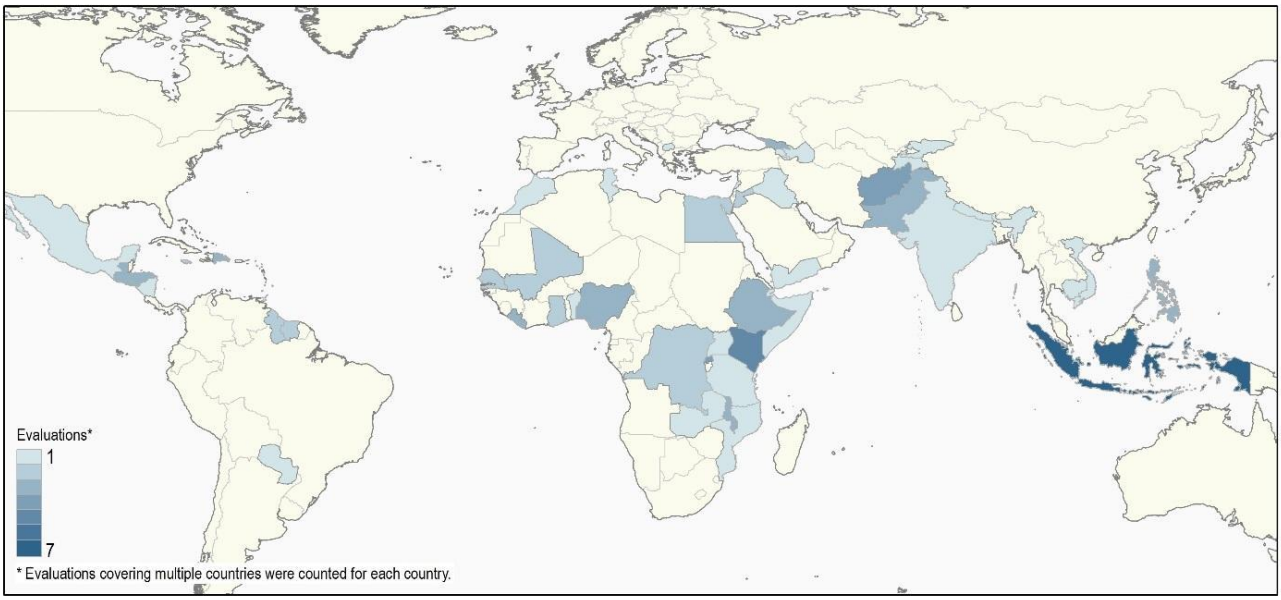
FINDINGS

The results described in this section are based on information captured using two tools. Technical experts used the evaluation quality tool mostly to capture expert judgments about elements of the evaluations. Annex 5 includes the full results.[30]

Context

The study team captured the country where the evaluated project or activity took place. As shown in Figure 5, the countries with the most evaluations reviewed were Indonesia, Afghanistan, Kenya, and Rwanda.

FIGURE 5: REVIEWED EVALUATIONS BY COUNTRY OF FOCUS



The team also categorized the evaluations by region, crisis and conflict status, and the income level in the country where the evaluated project or activity took place. The categories for region were based on the USAID standard classification. The Office of Education provided the study team with a list of countries that qualified as in crisis and conflict during the time span of the evaluations reviewed (2013–

[30] Annex 6 includes additional disaggregation of results from these reviews. MSI team members used an evaluation background tool mostly to capture basic contextual information about the evaluations.

2016). The categories for income were based on the World Bank's classification of countries' income levels. As shown in Figure 6, Africa accounted for the most evaluations, nearly two-thirds of the evaluated projects or activities were not in countries in crisis or conflict, and half of the evaluations examined projects or activities implemented in lower-middle income countries, while about one in five evaluations focused on an upper-middle income country.

FIGURE 6: PERCENTAGE OF EVALUATIONS BY REGION, CRISIS AND CONFLICT STATUS, AND COUNTRY INCOME

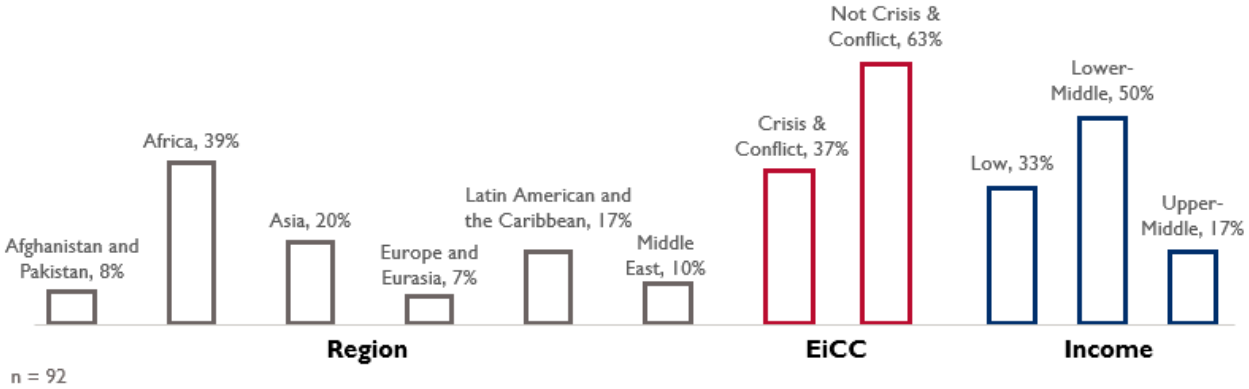


Figure 7 shows the report's year of publication and the phase of assessment. The study included 20 to 28 reports from each year and most were final evaluations (57 percent).

FIGURE 7: PERCENTAGE OF EVALUATIONS BY ASSESSMENT AND IMPLEMENTATION PHASES

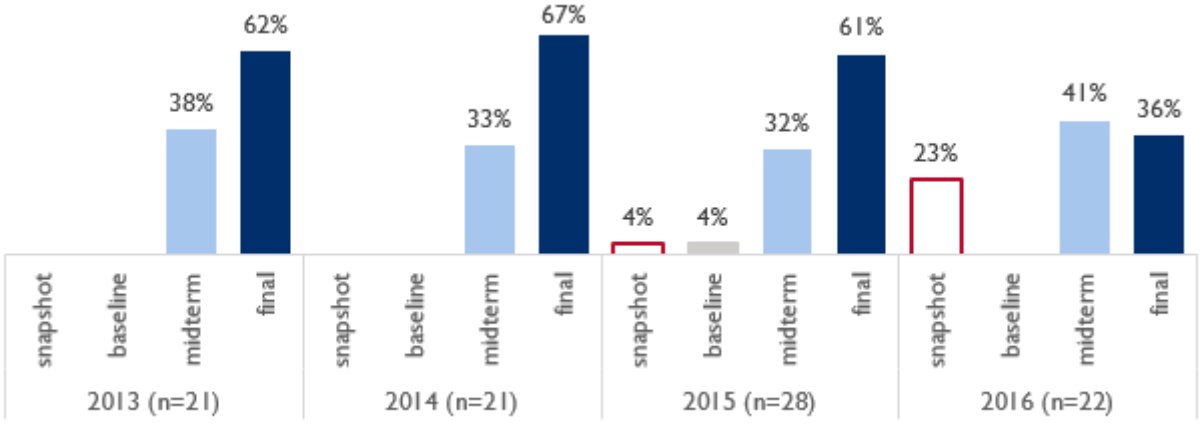
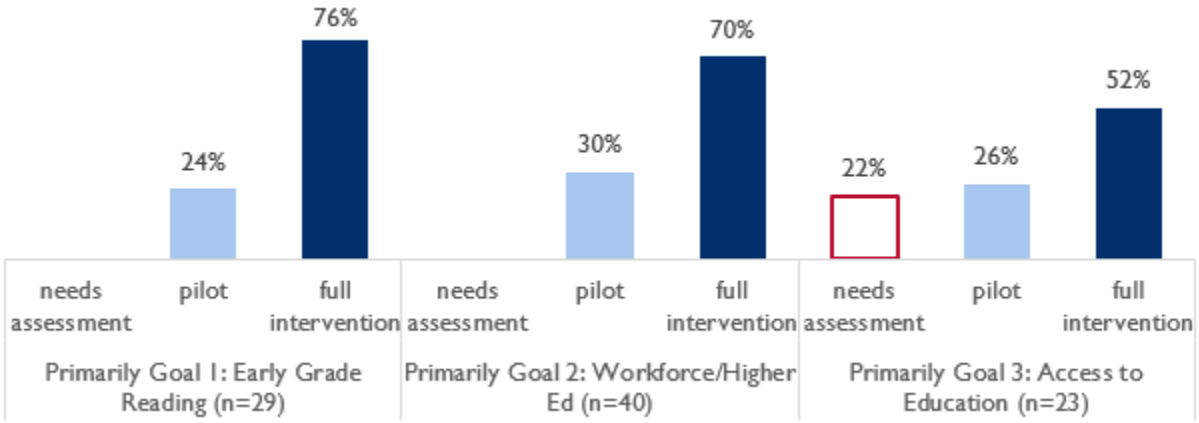


Figure 8 shows the primary Education Strategy Goal associated with each evaluation and the phase of implementation.[31] The study included 23 to 40 reports associated with each Education Strategy Goal and most evaluated full interventions (67 percent).

[31] For this study, the Office of Education instructed the study team to categorize evaluations to Education Strategy Goal 3 thematically as access to education instead of geographically as crisis and conflict.

FIGURE 8: PRIMARY EDUCATION STRATEGY GOAL BY IMPLEMENTATION PHASE



Per the USAID Evaluation Policy, the study team categorized evaluations as impact evaluations if they measured changes in development outcomes that were attributable to an intervention, and performance evaluations if they measured what a particular project or activity had achieved, whether expected results were occurring, how it was being implemented, or how it was perceived and valued. At the Office of Education's request, research studies that did not evaluate specific interventions were also included.

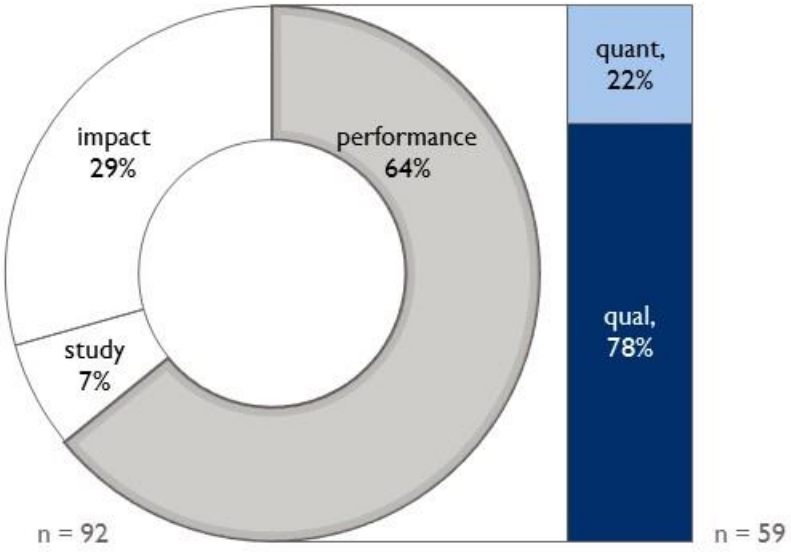
For this study:

- **Impact evaluations** were sub-categorized by having an experimental study design with a strong counterfactual, and the team recorded whether a cost-effectiveness analysis (CEA) complemented each impact evaluation.
- **Performance evaluations** were categorized as quantitative focused if they mostly collected quantitative data to track outcome measures, or qualitative focused if they mostly collected qualitative data about processes, such as how a project or activity was being implemented or how it was perceived and valued. For qualitative-focused performance evaluations, the team also recorded whether the report mentioned the use of an organizational capacity assessment (OCA).
- At the request of the Office of Education, the study included six **research studies** that did not evaluate a USAID-funded intervention. These were mostly needs assessments, and the team recorded whether the research study used a rapid education and risk analysis (RERA).

As shown in Figure 9, 64 percent of the evaluations reviewed were performance evaluations, and 29 percent were impact evaluations. Among the impact evaluations, 48 percent used an experimental design, and 19 percent included a CEA.[32] Among the performance evaluations, 78 percent were qualitative-focused process evaluations and among these only 13 percent used an OCA. Five of the six research studies included in this study were needs assessments, and four of the studies used a RERA.

[32] Evaluations were categorized as impact evaluations if described as such in the evaluation report. In the evaluation quality assessment, the expert reviewers then assessed whether the counterfactual used met standards of rigor.

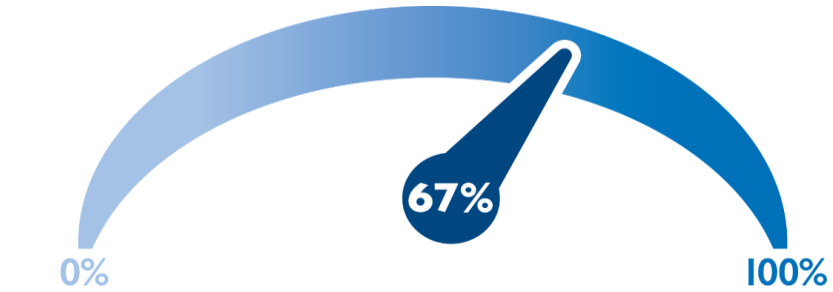
FIGURE 9: PERCENTAGE OF EVALUATIONS BY EVALUATION TYPE



Conceptual Framing

Conceptual framing included items and expert judgements related to the evaluation theory of change. Out of the seven principles of quality, conceptual framing had the second highest percentage of evaluations considered adequate.

FIGURE 10: PERCENTAGE OF EVALUATIONS RATED WITH ADEQUATE CONCEPTUAL FRAMING (N=92)



– ILLUSTRATION OF A NEGATIVE REVIEW

"The problem statement is not as clear as it could be. (…) There is no clear connection between the problem and the intervention. Next, it is difficult to understand the theory of change behind the intervention and how the ultimate impacts relate to the intermediate outcomes because there is no log frame or results framework."

+ ILLUSTRATION OF A POSITIVE REVIEW

"The research questions are clearly stated and directly tied to the project goals and objectives. The report would have benefited from additional contextual information but overall, the conceptual framing is adequate."

Figure 11 shows the percentage of evaluations scored as adequate in terms of conceptual framing by evaluation type, income level of the country where the evaluated project or activity was implemented, whether that project or activity was implemented in a conflict or crisis environment, and the primary Education Strategy Goal associated with the evaluation.

FIGURE 11: PERCENTAGE OF EVALUATIONS WITH ADEQUATE CONCEPTUAL FRAMING BY FACTOR

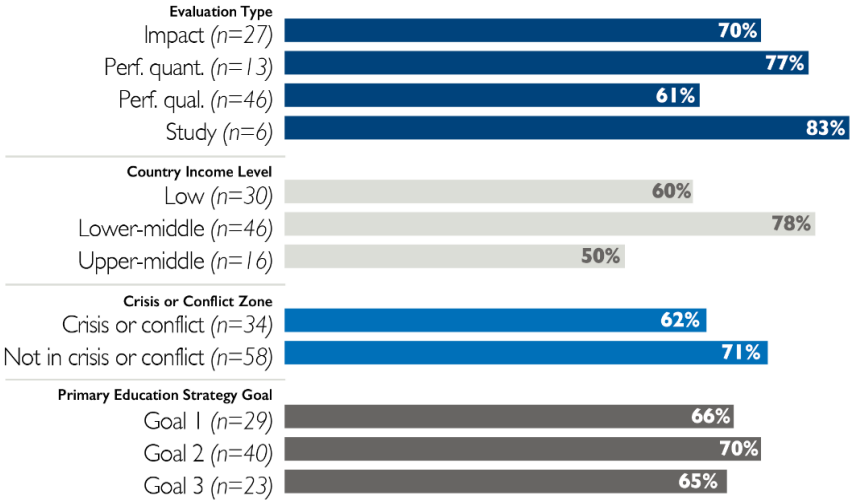


Figure 12 shows responses to the questions mapped to this principle for all evaluations. For two questions, "not applicable" was a possible response, but "not applicable" responses are not included in the percentages below.

FIGURE 12: PERCENTAGE OF EVALUATIONS BY CONCEPTUAL FRAMING ITEMS

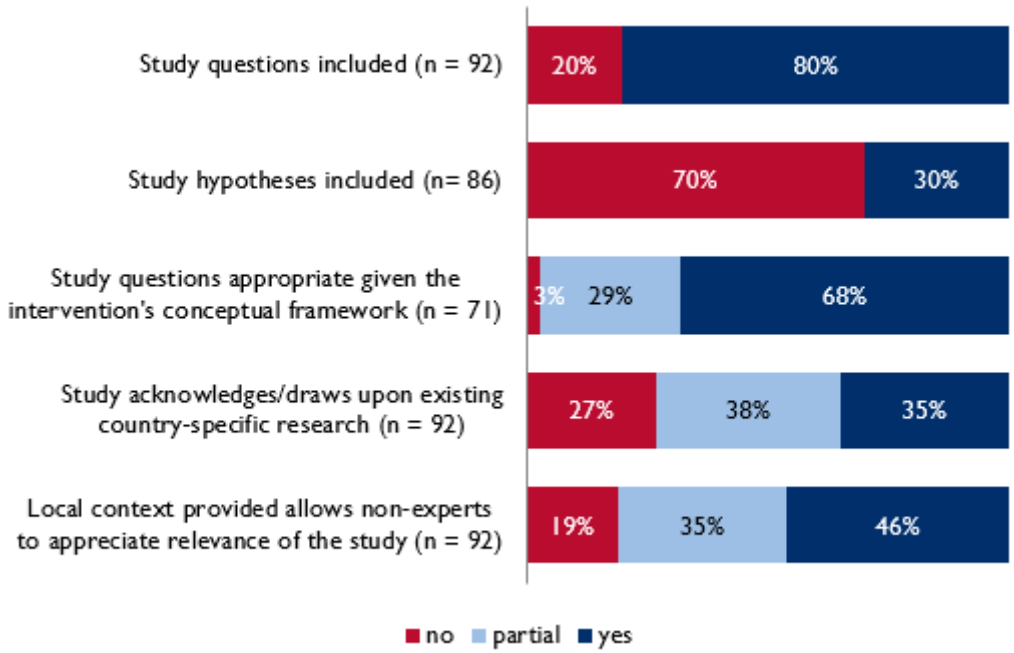
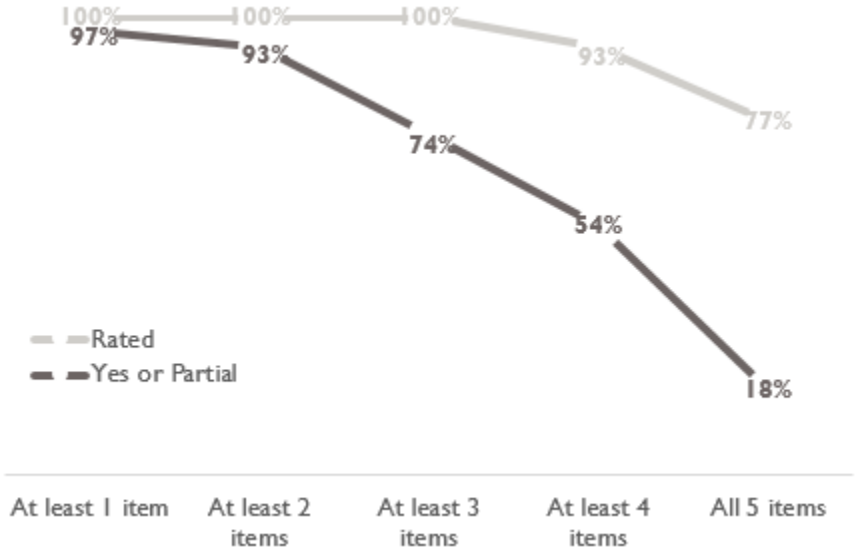


Figure 13 shows the percentage of evaluations by the number of questions rated and number of "yes" or "partial" responses.

FIGURE 13: PERCENTAGE OF EVALUATIONS BY CONCEPTUAL FRAMING ITEMS RATED AND FULLY OR PARTIALLY SATISFIED



Openness and Transparency

Openness and transparency included items and expert judgements related to the evaluation self-criticism and independence. Out of the seven principles of quality, openness and transparency had the fourth highest percentage of evaluations considered adequate.

FIGURE 14: PERCENTAGE OF EVALUATIONS WITH ADEQUATE OPENNESS AND TRANSPARENCY (N=92)

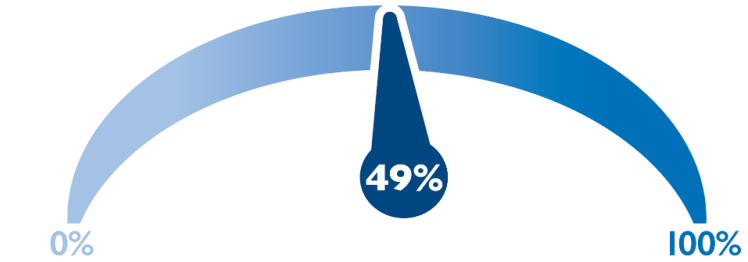


ILLUSTRATION OF A NEGATIVE REVIEW

"The report does not deal with the problems and decisions associated with sampling (why select control schools from those that did not apply to the program)…It did not report that the baseline and endline data collection team compositions may have differed, as it appears the implementing partner conducted the former and the evaluator the latter…Not transparent about team composition or limitations."

ILLUSTRATION OF A POSITIVE REVIEW

"The report presents limitations due to the timing of the evaluation, the availability of required data to address thoroughly each aspect of the evaluation framework and limitations due to the implementation of the projects."

Figure 15 shows the percentage of evaluations scored as adequate in terms of openness and transparency by evaluation type, income level of the country where the evaluated project or activity was implemented, whether that project or activity was implemented in a conflict or crisis environment, and the primary Education Strategy Goal associated with the evaluation.

FIGURE 15: PERCENTAGE OF EVALUATIONS WITH ADEQUATE OPENNESS AND TRANSPARENCY BY FACTOR

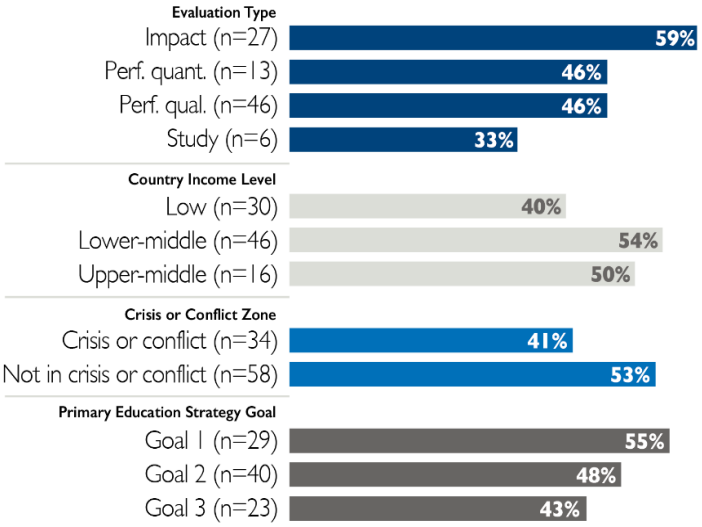


Figure 16 shows responses to the questions mapped to this principle for all evaluations. For two questions, "not applicable" was a possible response, but "not applicable" responses are not included in the percentages below.

FIGURE 16: PERCENTAGE OF EVALUATIONS BY OPENNESS AND TRANSPARENCY ITEMS

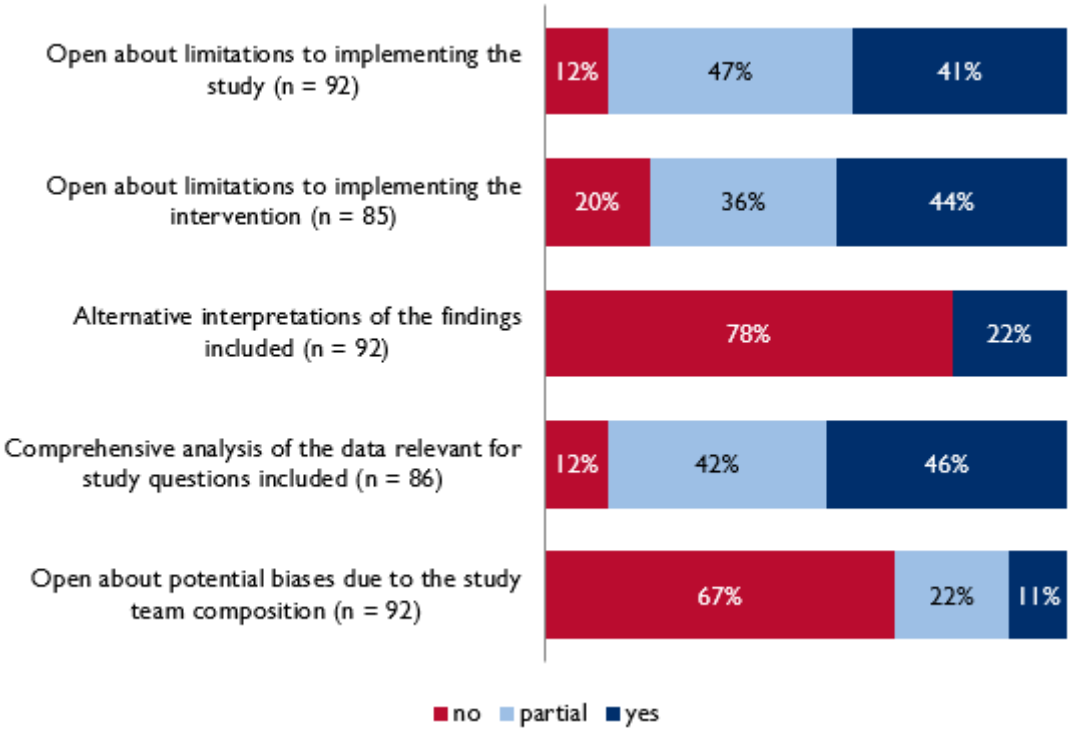
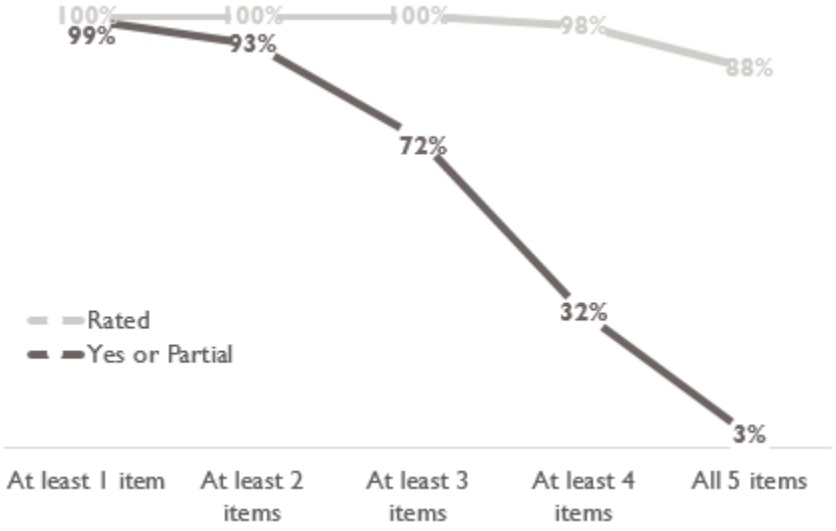


Figure 17 shows the percentage of evaluations by the number of questions rated and number of "yes" or "partial" responses.

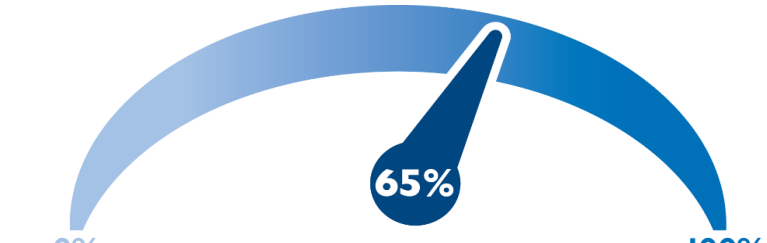
FIGURE 17: PERCENTAGE OF EVALUATIONS BY OPENNESS AND TRANSPARENCY ITEMS RATED AND FULLY OR PARTIALLY SATISFIED



Robustness of the Methodology

Robustness of the methodology included items and expert judgements related to the evaluation appropriateness of the design and rigorous implementation. Out of the seven principles of quality, robustness of the methodology had the third highest percentage of evaluations considered adequate.

FIGURE 18: PERCENTAGE OF EVALUATIONS WITH ADEQUATE ROBUSTNESS OF THE METHODOLOGY (N=92)



— ILLUSTRATION OF A NEGATIVE REVIEW

"There was almost no triangulation of data, and we don't know how/why these [number] schools were selected out of the [number] implementing, so we don't really know how much confidence we should have in the results. On the plus side, the authors were clear that there was to be a more rigorous program evaluation with an experimental design, but that was not possible so they did what they could to learn about the program implementation and effects."

+ ILLUSTRATION OF A POSITIVE REVIEW

"The authors adopted a methodology that is appropriate for answering all the posed research questions. The evaluation matrix they elaborated allows the readers to have a clear picture about the methods used to answer each of the questions. It also highlighted the data triangulation that was conceived as part of the methodology."

Figure 19 shows the percentage of evaluations scored as adequate in terms of robustness of the methodology by evaluation type, income level of the country where the evaluated project or activity was implemented, whether that project or activity was implemented in a conflict or crisis environment, and the primary Education Strategy Goal associated with the evaluation.

FIGURE 19: PERCENTAGE OF EVALUATIONS WITH ADEQUATE ROBUSTNESS OF THE METHODOLOGY BY FACTOR

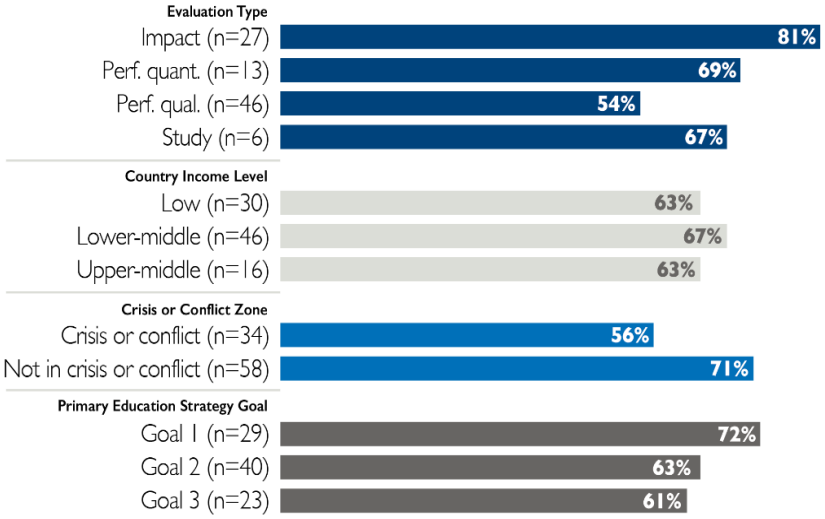


Figure 20 shows responses to the questions mapped to this principle for all evaluations. For two questions, "not applicable" was a possible response, but "not applicable" responses are not included in the percentages below.

FIGURE 20: PERCENTAGE OF EVALUATIONS BY ROBUSTNESS OF THE METHODOLOGY ITEMS

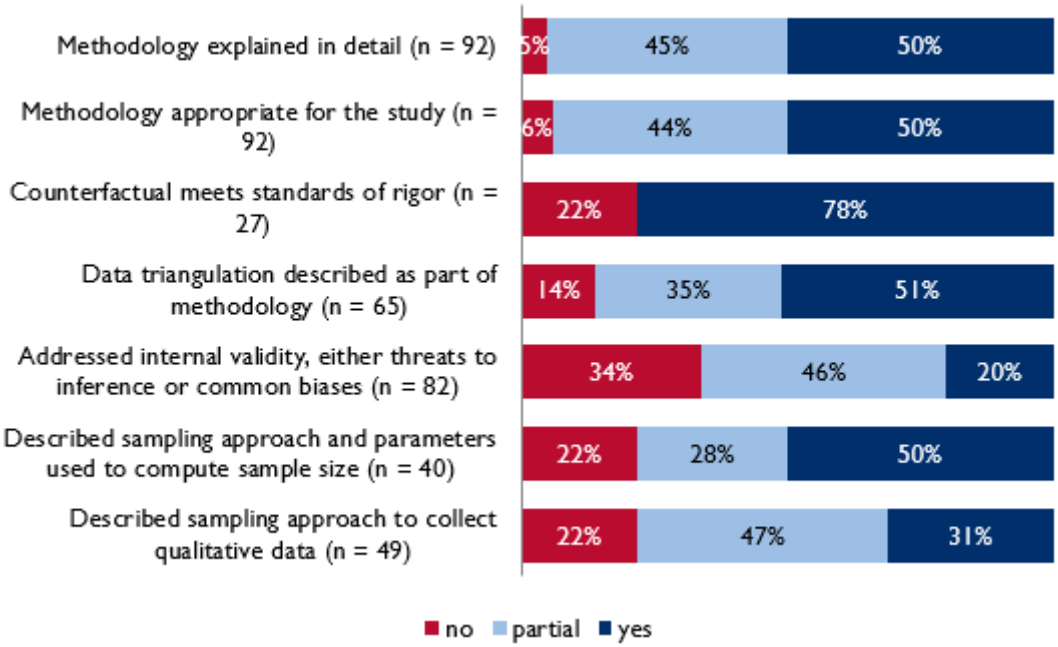
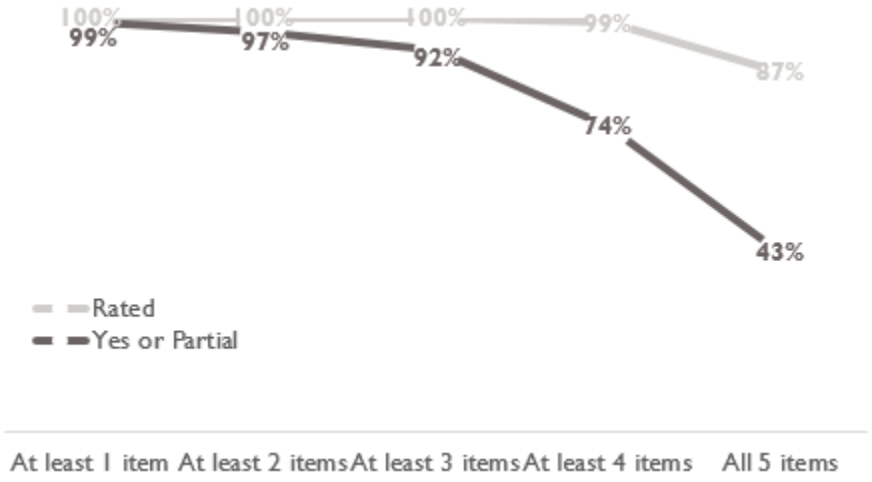


Figure 21 shows the percentage of evaluations by the number of questions rated and number of "yes" or "partial" responses.[33]

FIGURE 21: PERCENTAGE OF EVALUATIONS BY ROBUSTNESS OF THE METHODOLOGY ITEMS RATED AND FULLY OR PARTIALLY SATISFIED



[33] While seven items were matched to this principle, two were variations of a similar construct tailored to specific evaluation types. Therefore, up to five items were applied to an evaluation.

Cultural Appropriateness

Cultural appropriateness included items and expert judgements related to the evaluation's use of culturally relevant tools and culturally sensitive analysis. Out of the seven principles of quality, cultural appropriateness had the lowest percentage of evaluations considered adequate.

FIGURE 22: PERCENTAGE OF EVALUATIONS WITH ADEQUATE CULTURAL APPROPRIATENESS (N=92)

Figure 23 shows the percentage of evaluations scored as adequate in terms of cultural appropriateness by evaluation type, income level of the country where the evaluated project or activity was implemented, whether that project or activity was implemented in a conflict or crisis environment, and the primary Education Strategy Goal associated with the evaluation.

FIGURE 23: PERCENTAGE OF EVALUATIONS WITH ADEQUATE CULTURAL APPROPRIATENESS BY FACTOR

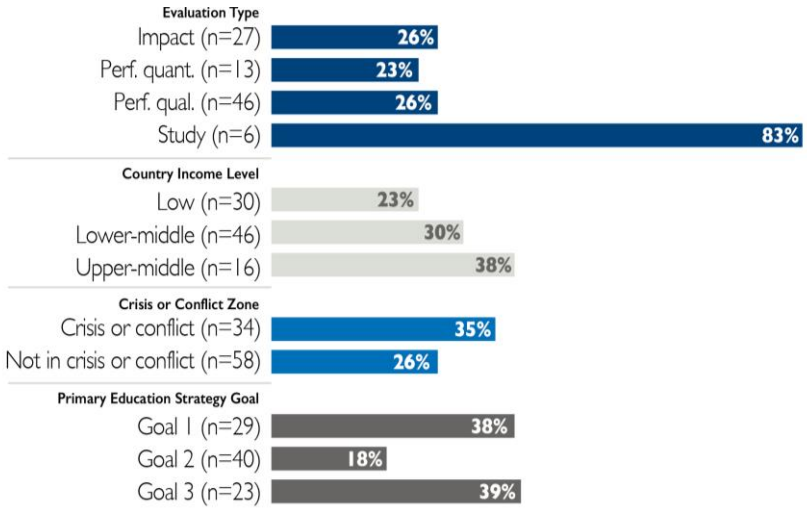


Figure 24 shows responses to the questions mapped to this principle for all evaluations. For two questions, "not applicable" was a possible response, but "not applicable" responses are not included in the percentages below.

FIGURE 24: PERCENTAGE OF EVALUATIONS BY CULTURAL APPROPRIATENESS ITEMS

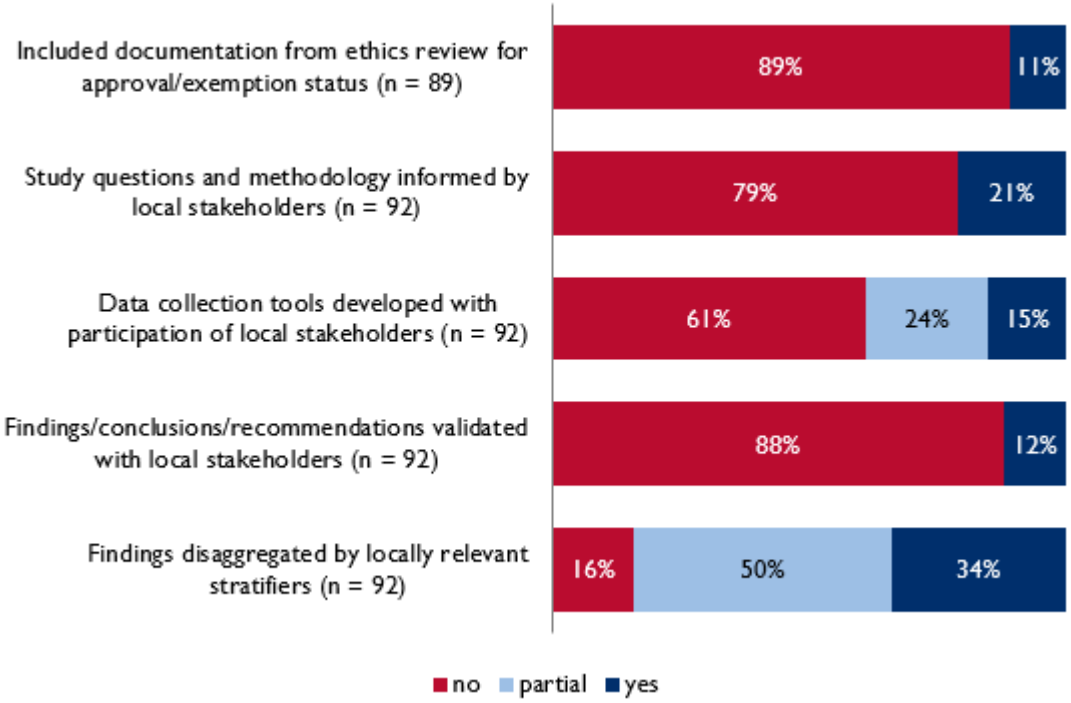
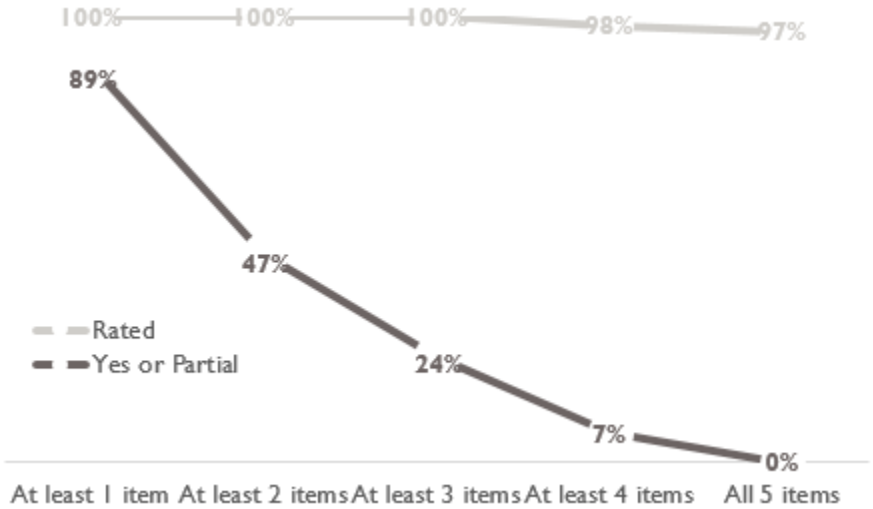


Figure 25 shows the percentage of evaluations by the number of questions rated and number of "yes" or "partial" responses.

FIGURE 25: PERCENTAGE OF EVALUATIONS BY CULTURAL APPROPRIATENESS ITEMS RATED AND FULLY OR PARTIALLY SATISFIED



Validity

Validity included items and expert judgements related to the evaluation measurement, internal, external, and ecological validity. Out of the seven principles of quality, validity had the fifth highest percentage of evaluations considered adequate.

FIGURE 26: PERCENTAGE OF EVALUATIONS WITH ADEQUATE VALIDITY (N=92)

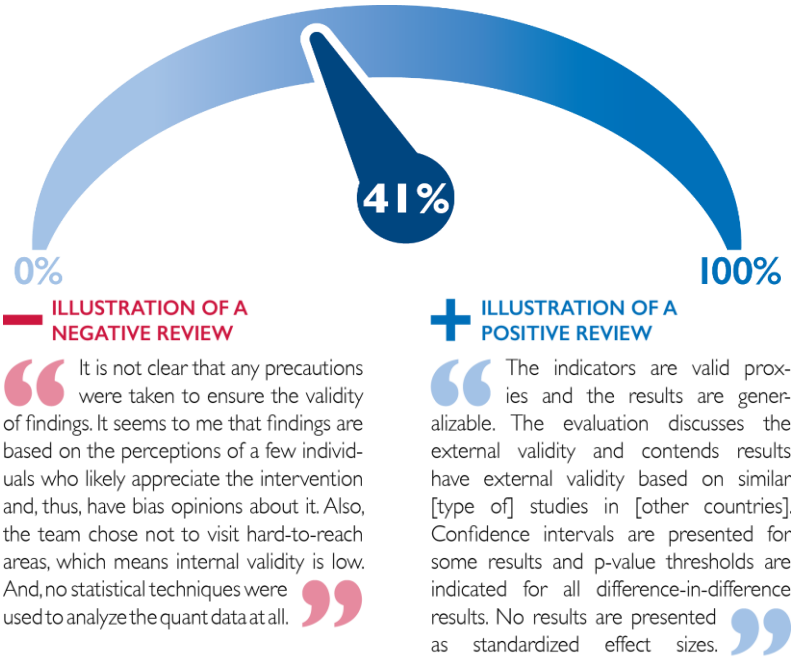


Figure 27 shows the percentage of evaluations scored as adequate in terms of validity by evaluation type, income level of the country where the evaluated project or activity was implemented, whether that project or activity was implemented in a conflict or crisis environment, and the primary Education Strategy Goal associated with the evaluation.

FIGURE 27: PERCENTAGE OF EVALUATIONS WITH ADEQUATE VALIDITY BY FACTOR

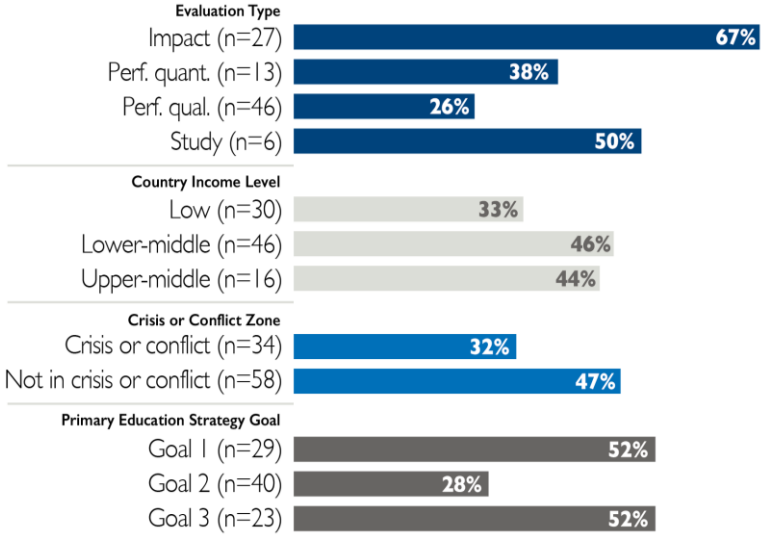


Figure 28 shows responses to the questions mapped to this principle for all evaluations. For two questions, "not applicable" was a possible response, but "not applicable" responses are not included in the percentages below.

FIGURE 28: PERCENTAGE OF EVALUATIONS BY VALIDITY ITEMS

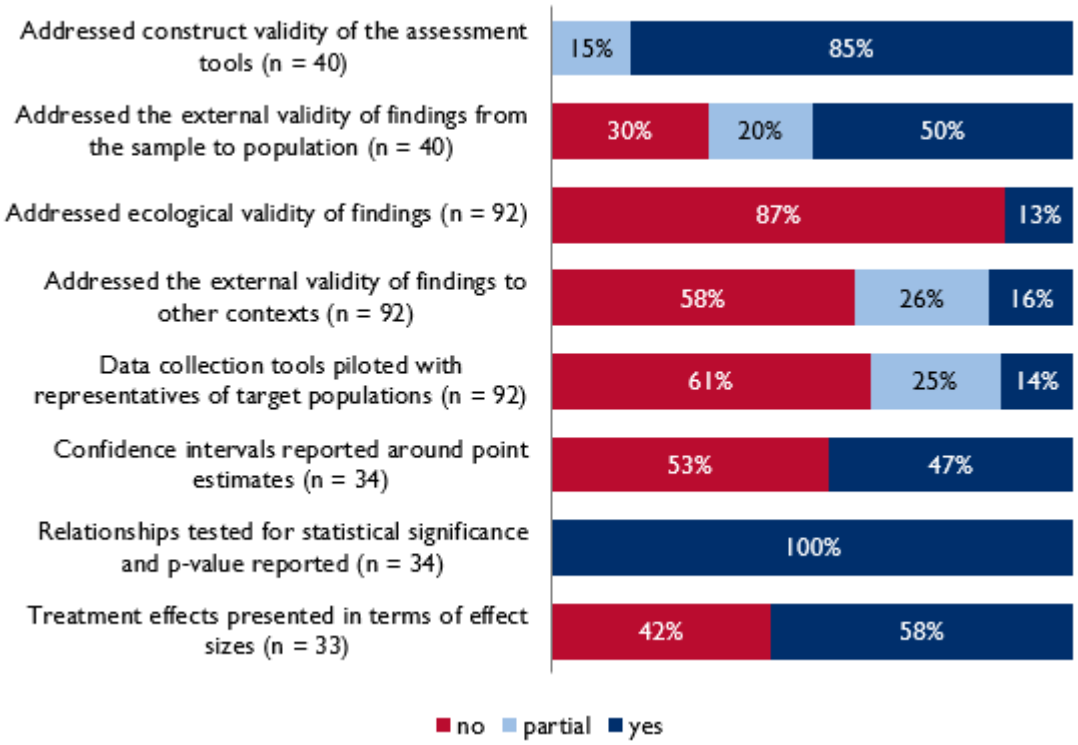
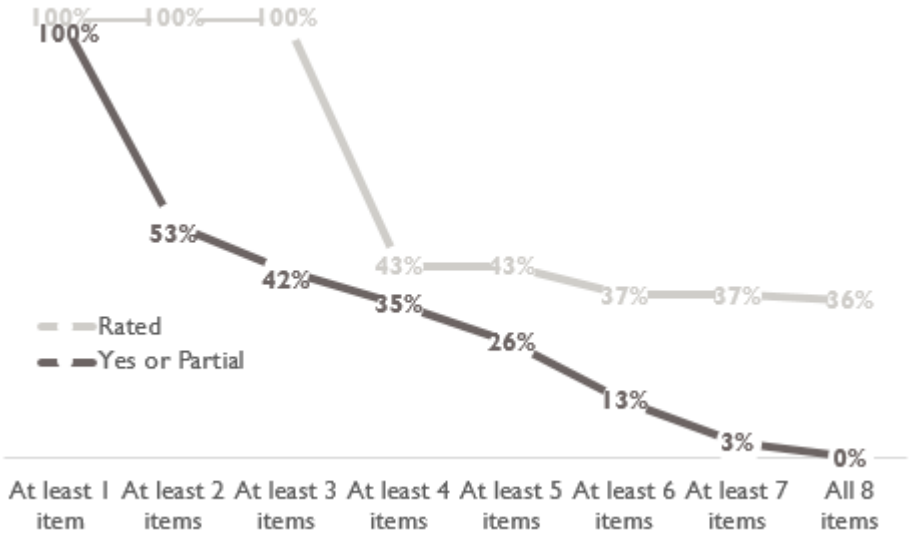


Figure 29 shows the percentage of evaluations by the number of questions rated and number of "yes" or "partial" responses.

FIGURE 29: PERCENTAGE OF EVALUATIONS BY VALIDITY ITEMS RATED AND FULLY OR PARTIALLY SATISFIED



Reliability

Reliability included items and expert judgements related to the evaluation consistent measurement and consistent results from repeated processing and analysis. Out of the seven principles of quality, reliability had the sixth highest percentage of evaluations considered adequate.

FIGURE 30: PERCENTAGE OF EVALUATIONS WITH ADEQUATE RELIABILITY (N=92)

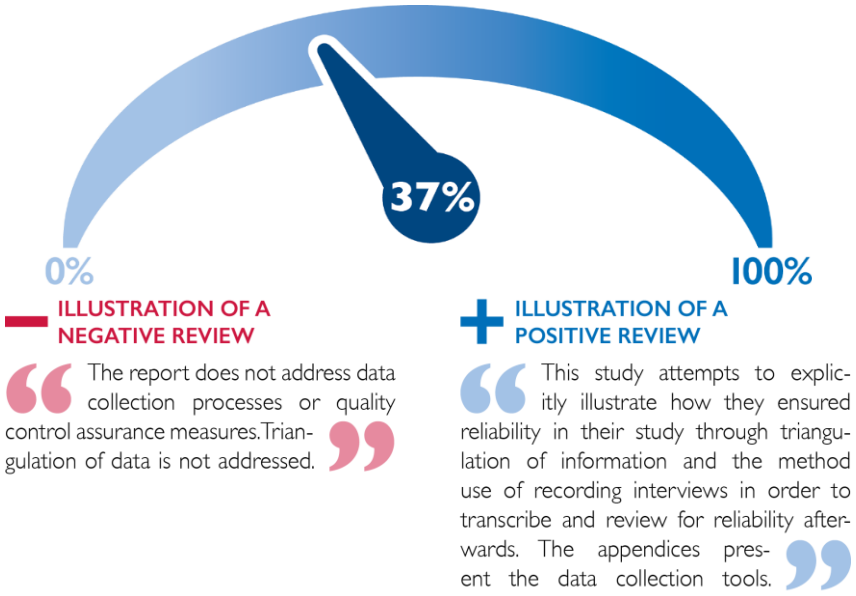


Figure 31 shows the percentage of evaluations scored as adequate in terms of reliability by evaluation type, income level of the country where the evaluated project or activity was implemented, whether that project or activity was implemented in a conflict or crisis environment, and the primary Education Strategy Goal associated with the evaluation.

FIGURE 31: PERCENTAGE OF EVALUATIONS WITH ADEQUATE RELIABILITY BY FACTOR

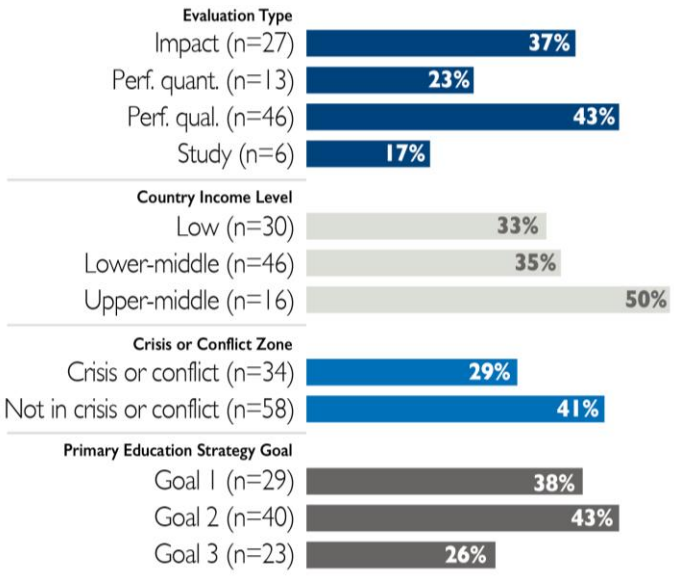


Figure 32 shows responses to the questions mapped to this principle for all evaluations. For two questions, "not applicable" was a possible response, but "not applicable" responses are not included in the percentages below.

FIGURE 32: PERCENTAGE OF EVALUATIONS BY RELIABILITY ITEMS[34]

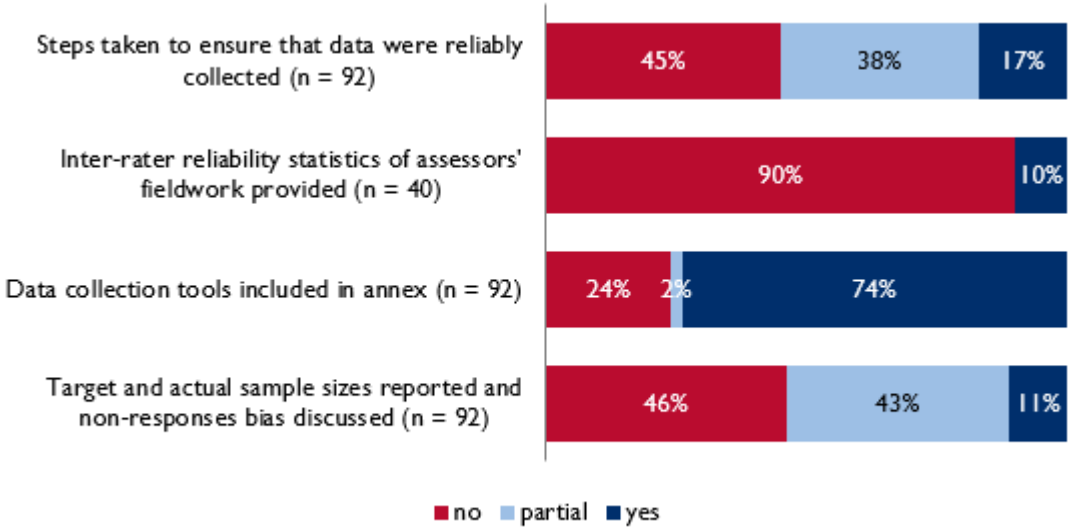
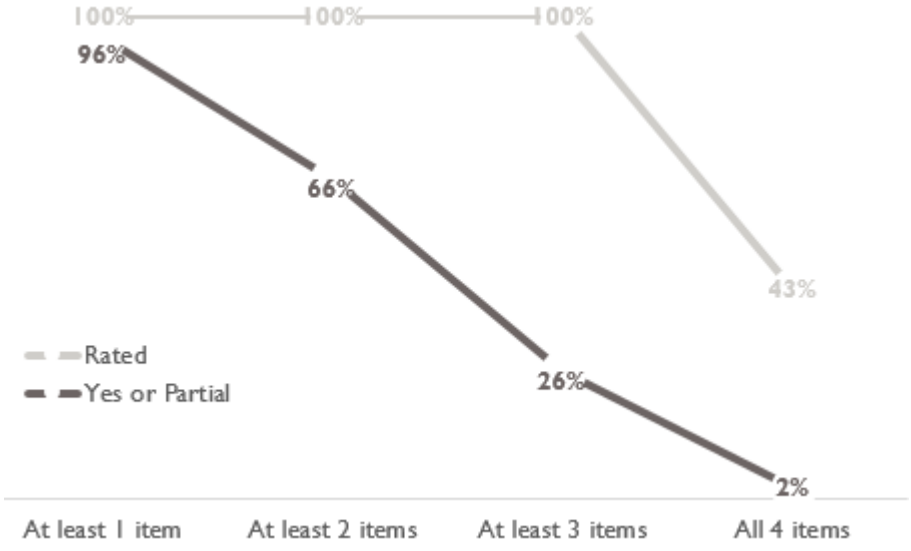


Figure 33 shows the percentage of evaluations by the number of questions rated and number of "yes" or "partial" responses.

FIGURE 33: PERCENTAGE OF EVALUATIONS BY RELIABILITY ITEMS RATED AND FULLY OR PARTIALLY SATISFIED



[34] USAID does not require the inclusion of inter-rater reliability results in reports. As such, it is possible that inter-rater reliability has been collected but not discussed in the evaluation report.

Cogency

Cogency included items and expert judgements related to the evaluation logical argumentative thread throughout the report and conclusions being based on the evaluation's findings. Out of the seven principles of quality, conceptual framing had the highest percentage of evaluations considered adequate.

FIGURE 34: PERCENTAGE OF EVALUATIONS WITH ADEQUATE COGENCY (N=92)

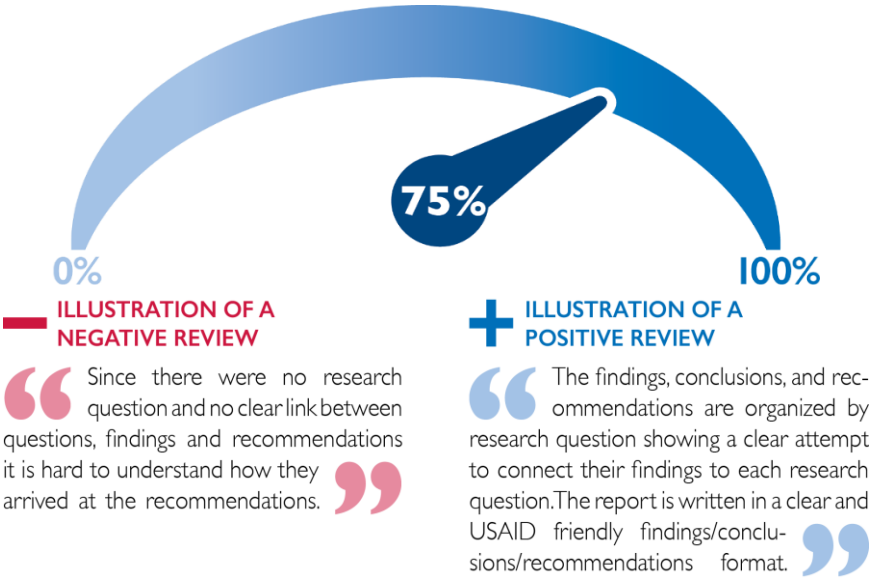


Figure 35 shows the percentage of evaluations scored as adequate in terms of cogency by evaluation type, income level of the country where the evaluated project or activity was implemented, whether that project or activity was implemented in a conflict or crisis environment, and the primary Education Strategy Goal associated with the evaluation.

FIGURE 35: PERCENTAGE OF EVALUATIONS WITH ADEQUATE COGENCY BY FACTOR

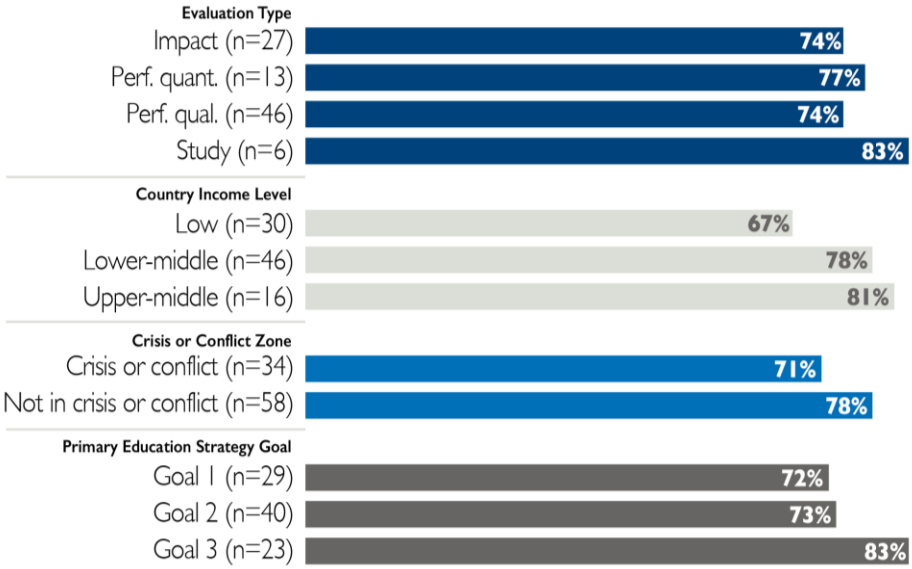


Figure 36 shows responses to the questions mapped to this principle for all evaluations. For two questions, "not applicable" was a possible response, but "not applicable" responses are not included in the percentages below.

FIGURE 36: PERCENTAGE OF EVALUATIONS BY COGENCY ITEMS

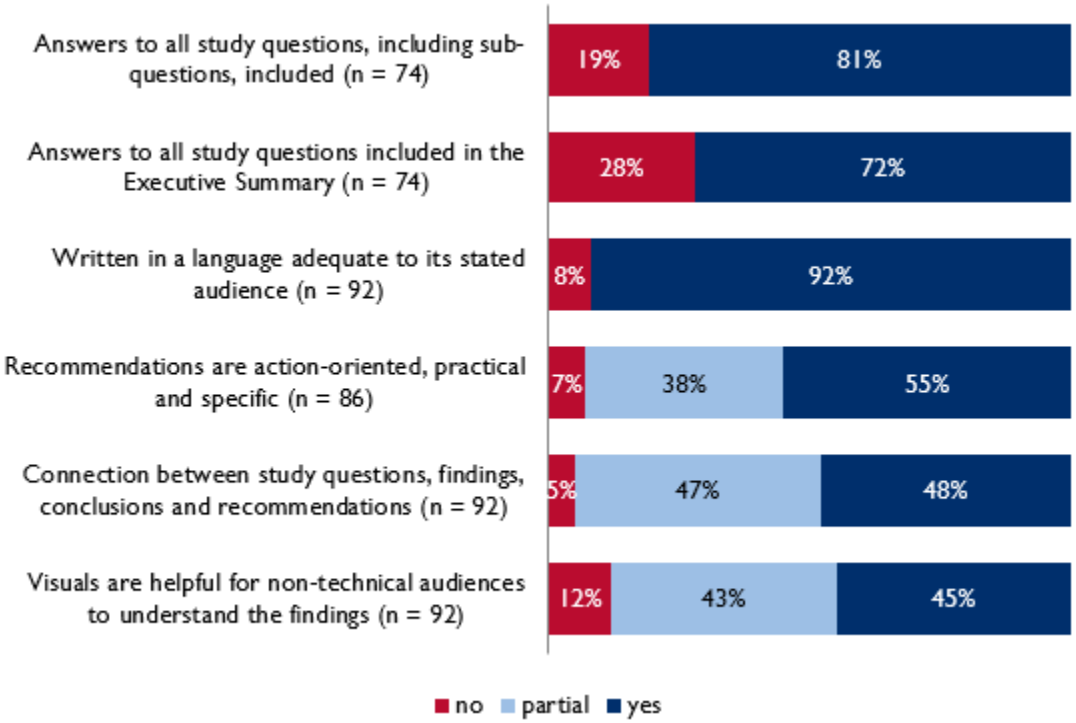
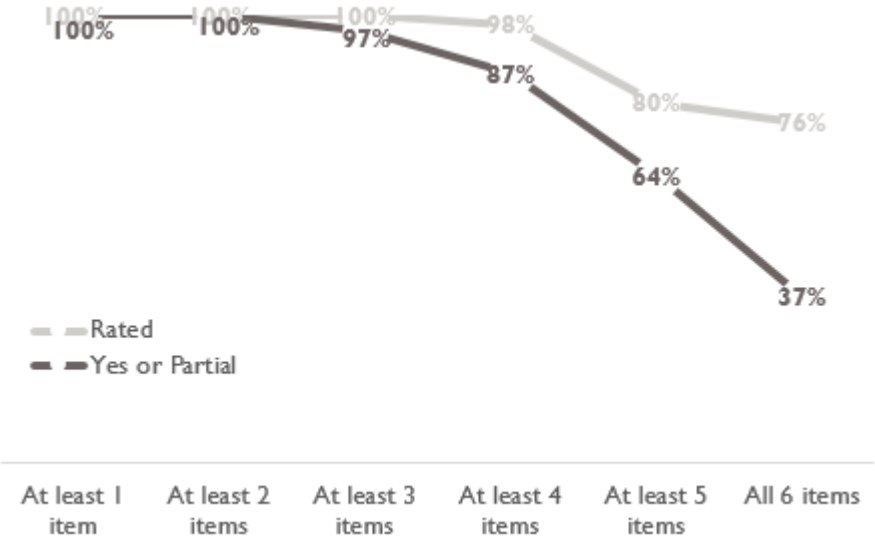


Figure 37 shows the percentage of evaluations by the number of questions rated and number of "yes" or "partial" responses.

FIGURE 37: PERCENTAGE OF EVALUATIONS BY COGENCY ITEMS ATTEMPTED AND FULLY OR PARTIALLY SATISFIED



DISCUSSION

Development of Evaluation Quality Tool: Lessons Learned

Reviewers' overall feedback on the evaluation quality tool was positive and constructive, aimed at improving the tool. Reviewers voiced interest in participating in a future round of evaluation quality reviews, which aligns with the Office of Education's interest in repeating evaluation quality reviews on a periodic basis. They agreed that the tool should not be used to produce a composite score about overall evaluation quality, and that the scoring of the adequacy of the principles of quality should be relative to the evaluation type. They also mentioned other circumstances to consider, such as adding value for money. The BE^2 framework considers value for money a desired, if not mandatory, dimension, and the GAO performance audit captured it; however, this was beyond the scope of the present study. As more USAID-funded evaluations in the education sector include cost-effectiveness analysis, value for money could be included as a principle in future evaluation quality reviews.

Much of the additional feedback provided during the full-day reviewers' meeting related to the constraints that implementers and evaluators face. While USAID's guidance has been helpful for both implementing and evaluation partners, it focuses on supply-side aspects of the evaluations. USAID may wish to consider the demand side as well. Reviewers noted that in addition to best practices and USAID guidance, evaluations need to be responsive to a statement of work. Thus, ensuring that these statements of work are technically sound and based on a broader conceptual framework may help evaluation partners to further support learning objectives. Suggestions from the reviewers' meeting included having joint trainings between USAID education officers and implementing and evaluation partners on key topics.

As mentioned in BE^2 Assessing the Strength of Evidence in the Education Sector, it has been increasingly recognized that mixed methods designs that use sequential data collection should be used to bolster a study's exploratory or explanatory power.[35] As these designs may become more prevalent in the future, the evaluation quality tool might need to be adapted to accommodate this development.

Assessment of the Quality of Evaluations: Lessons Learned

The consensus ratings of the co-reviewers of each evaluation provide useful insights into areas of strength in current practice and areas that might be lacking. Most evaluations satisfied basic reporting requirements.[36] This aligns with previous assessments of the compliance of evaluation reports across E3

[35] Explanatory Design: Collection and analysis of quantitative data followed by the subsequent collection and analysis of qualitative data. The qualitative phase of the study follows from the results of the quantitative phase. For instance, starting with a quantitative survey study, one identifies statistically significant differences and anomalous results, then follows up these results with an in-depth qualitative study to explain why these results occurred. Exploratory Design: Collection and analysis of qualitative data followed by collection and analysis of quantitative data. The quantitative phase of this study builds on the exploration of the phenomenon done in the qualitative phase. For instance, starting with an in-depth qualitative study, one gathers the information necessary for developing an instrument, identifying variables, or stating propositions for testing, then follows up by using this information in a quantitative study. For more: John W. Creswell and Vicki L. Plano Clark, *Designing and Conducting Mixed Methods Research* (California: SAGE Publications, 2011).

[36] For instance, four of five evaluations in this study included the study questions, three of four evaluations included data collection instruments, and one of two evaluations that used inferential statistical methods reported confidence intervals. Exceptions, such as 9 of 10 evaluations not providing inter-rater reliability statistics for assessor's fieldwork, may be a function of the newness or selectivity of the requirement.

sectors, which showed that evaluations in the education sector tended to outperform evaluations for the other E3 sectors in this regard.[37]

Overall, evaluations reviewed showed greater strength in cogency, conceptual framing, robustness of methodology, and openness and transparency, and greater weakness in validity, reliability, and cultural appropriateness. These results align with findings from the GAO performance audit, Agencies Can Improve the Quality and Dissemination of Program Evaluations, which reviewed final reports for 49 performance and 14 impact evaluations from all sectors within USAID. This audit recorded the percentage of evaluations that generally met a similar set of quality criteria. The criteria met by the highest percentage of evaluations were: the study questions align with the key stated goals of the intervention (conceptual framing); the evaluation design is appropriate given the study questions (robustness of the methodology); and the conclusions are supported by the available evidence (cogency). The criteria met by the lowest percentage of evaluations were: data collection is appropriate and data analysis appears appropriate (reliability); and the target population and sampling for the evaluation are appropriate (validity).[38]

Thus, while this study and the GAO performance audit differ in other respects, including the timing, stage, and sectors of the evaluations reviewed, the similar order in which the principles of quality are ranked in these two studies suggests that strengthening the validity and reliability of evaluations is a challenge at the office and Agency levels. The findings from the GAO performance audit, which extended beyond USAID, suggest that this is also a struggle for other foreign assistance agencies, which may speak to the difficulties of evaluating programs abroad in often challenging environments.

The study team also examined patterns in the reviews by the income level of the country where the evaluation took place, whether that project or activity was implemented in a crisis and conflict environment, the primary Education Strategy Goal associated with the evaluation, and the evaluation type. Neither income level nor crisis and conflict status was strongly associated with the consensus responses to the evaluation quality assessment. This implies that the **country's income level and crisis and conflict status are poor predictors of quality for the USAID-funded evaluations in education.**

Only two items were strongly associated with country income level. Whether the evaluation addressed the external validity of findings from the sample to population showed a negative trend with country income level (Cramer's V = 0.3003), with the percentage of evaluations failing to address this item increasing from low to lower-middle to upper-middle income countries (14, 38, and 50 percent, respectively). The team found a similar trend for whether the treatment effects were presented in terms of effect sizes. However, both items were applicable only to impact and outcome performance evaluations, which were seldom in the upper-middle income countries included in this study.

[37] The Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009-2012, based on 340 evaluations (27 related to the education sector), found an average evaluation report quality score for education sector evaluations of 6.5, compared to 5.9 for the E3 Bureau overall. Similarly, the Sectoral Synthesis of 2013-2014 Evaluation Findings, based on 117 evaluations (42 related to the education sector), found an average evaluation report quality score for education sector evaluations of 8.2, compared to 8.0 for the E3 Bureau overall. The Sectoral Synthesis of FY2015 Evaluation Findings, based on 92 evaluations (29 related to the education sector), found an average evaluation report quality score for education sector evaluations of 7.9, compared to 7.5 for the E3 Bureau overall.

[38] The GAO quality criteria did not directly address openness and transparency or cultural appropriateness. Two of the GAO quality criteria mapped somewhat to items under the principles of quality in this study's Evaluation Quality tool. The GAO criterion "the chosen indicators are appropriate for the study objectives" partially maps to an item about construct validity under the validity principle in this study, and similarly high percentages of evaluations met the criterion in both studies. The GAO quality criterion "the recommendations and lessons learned are supported by the available evidence" maps to an item about connections between findings, conclusions, and recommendations under the cogency principle in this study, and similarly low percentages of evaluations met the criterion in both studies.

The primary Education Strategy Goal for each evaluation also was not strongly associated with most of the consensus responses to the evaluation quality assessment. As previously mentioned, for this study the Office of Education instructed that the categorizing of evaluations to Education Strategy Goal 3 be defined thematically as access to education instead of geographically as crisis and conflict. Four items were strongly associated with the primary Education Strategy Goal of evaluations: whether the evaluation addressed the external validity of findings from the sample to population (Cramer's V = 0.4677), whether the counterfactual met standards of rigor (Cramer's V = 0.4245), whether the report included documentation from ethics review for approval/exemption status (Cramer's V = 0.3093), and whether the report included confidence intervals around point estimates (Cramer's V = 0.3234).

The frequency distribution indicates: (i) that a sizable proportion of impact evaluations associated with Goal 1 have counterfactuals that do not meet standards of rigor (42 percent), and a sizable proportion of impact and outcome performance evaluations associated with Goal 1 did not report confidence intervals (71 percent); (ii) that evaluations associated with Goal 2 lagged behind in terms of addressing the generalizability of findings from the sample to the population (27 percent), and USAID funded fewer impact evaluations relating to Goal 2 than the other goals; and (iii) that most evaluations did not include documentation from ethics review for approval/exemption status, with Goal 3 having the highest percentage of evaluations that reported ethics reviews (27 percent). **Notably, results indicate that the Education Strategy Goal is a poor predictor of whether the evaluation adequately addressed principles of quality.**

The evaluation type was often strongly associated with the consensus responses to the evaluation quality assessment,[39] with 11 items showing a strong association with evaluation type. The results suggest three classes of patterns: (1) a directional pattern, where responses tracked the continuum from more quantitative impact evaluations, to more quantitative performance evaluations, to more qualitative performance evaluations; (2) a pattern of extremes, where responses to the extremes of this continuum (fully quantitative or fully qualitative) were more similar than those in the middle; and (3) a more categorical pattern where responses tracked the distinction between performance and impact evaluations. This section discusses each of these in turn.

Directional pattern

Qualitative evaluations appeared to be more likely to include evaluation questions in the report, with percentages for impact evaluations, quantitative performance evaluations, and qualitative performance evaluations being 67, 85, and 91 percent, respectively (Cramer's V = 0.3359). In contrast, quantitative evaluations appeared to be more likely to include study hypotheses, with percentages for impact evaluations, quantitative performance evaluations, and qualitative performance evaluations being 56, 39, and 13 percent, respectively (Cramer's V = 0.3359). Quantitative evaluations also appeared more likely to be judged as adequate with respect to validity, where percentages for impact evaluations, quantitative performance evaluations, and qualitative performance evaluations were 67, 39, and 26 percent, respectively (Cramer's V = 0.3580).

One interpretation of these patterns is that the accountability purpose of evaluations (e.g., whether the project or activity worked) might often be assumed in more quantitative evaluations and therefore not explicitly included in the evaluation (or its scope of work), while the learning purpose of evaluations (e.g., which elements of the theory of change were validated by the project or activity's results) might be made more explicit, for example by setting a development hypothesis to be quantitatively tested. A corollary of this might be a greater existing emphasis on issues relating to validity with respect to more quantitative than qualitative evaluations. In turn, this might indicate that the Office of Education's

[39] For the sake of clarifying patterns across evaluation types, this section does not discuss results for research studies that did not evaluate a specific intervention or effects driven solely by such studies.

learning agenda could benefit from further consideration of recommendations for qualitative evaluations, especially in regard to leveraging their complementary exploratory and explanatory power to quantitative evaluations through sequential data collection in mixed methods evaluations.[40]

Pattern of extremes

Quantitative performance evaluations appeared less likely than other evaluation types to discuss alternative interpretations of the findings, with the proportion of impact evaluations, quantitative performance evaluations, and qualitative performance evaluations discussing alternative interpretations at 41, 8, and 13 percent, respectively (Cramer's V = 0.3256). Quantitative performance evaluations were similarly less open about limitations in implementing the intervention, with the proportion of impact evaluations, quantitative performance evaluations, and qualitative performance evaluations discussing limitations at 59, 8, and 44 percent, respectively (Cramer's V = 0.3256).

One tentative explanation is that if quantitative performance evaluations are primarily focused on tracking purposes, discussion of alternative interpretations and limitations might be seen as less of a priority.

Categorical pattern

Impact evaluations appeared substantially more likely than performance evaluations to include a comprehensive analysis of the data relevant for study questions, with 78 percent of impact evaluations, 31 percent of quantitative performance evaluations, and 40 percent of qualitative performance evaluations including such a comprehensive analysis (Cramer's V = 0.3662).

Performance evaluations were also more likely than impact evaluations to be judged as failing to address numerous components of validity (here a higher percentage is worse). The study team observed this kind of association for: internal validity (either threats to inference or common biases), which 8 percent of impact evaluations, 46 percent of quantitative performance evaluations, and 46 percent of qualitative performance evaluations failed to address (Cramer's V = 0.3231); (ii) external validity of findings to other contexts, which 37 percent of impact evaluations, 77 percent of quantitative performance evaluations, and 67 percent of qualitative performance evaluations failed to discuss (Cramer's V = 0.3570).

Items that were not applicable to qualitative performance evaluations but which showed a similar trend across the quantitative performance evaluations and impact evaluations were: construct validity of the assessment tools, which impact and quantitative performance evaluations failed to address at 7 and 31 percent, respectively (Cramer's V = 0.3064) and; external validity of findings from the sample to the population, which impact and quantitative performance evaluations failed to address at 11 and 69 percent, respectively (Cramer's V = 0.6062).

The only item for which a higher proportion of performance evaluations than impact evaluations scored positively was confidence intervals around point estimates, which 37 percent of impact evaluations and 86 percent of quantitative performance evaluations reported (Cramer's V = 0.3943). This might be because quantitative performance evaluations usually report the outcome of simple hypothesis testing for before and after comparisons, and there is a well-established tradition in USAID-funded evaluations of presenting the mean difference and associated confidence interval. Impact evaluations, by contrast,

[40] It has been suggested that most new empirical work on estimating the impact on learning of various education programs has been based on rigorous methods of estimating causal impacts; however, evaluations have inadequate conceptual framings, which jeopardizes their usefulness in formulating effective actions. As a result, "more of the same" empirical research might be unlikely to add up to a coherent research agenda. See Lant Pritchett, *The Evidence About What Works in Education: Graphs to Illustrate External Validity and Construct Validity* (RISE Insights, June 2017).

report the outcome of difference-in-differences estimates or more complex regression models and there might be less of an established practice in USAID-funded impact evaluations of providing the associated confidence interval.

Taken as a whole, these patterns might suggest that the emphasis that the Agency and other donors have put on improving the quality of impact evaluations has been successful in improving their validity, but that this positive outcome has not yet transferred to performance evaluations. This might be because a push in the international development community and USAID for rigorous testing of pilot interventions provided much guidance on how to design and implement impact evaluations, while guidance for designing and implementing rigorous performance evaluation remained more scarce.

CONCLUSION

This study demonstrated the benefits of assessing the quality of USAID-funded evaluations in the education sector using a holistic framework that maps different aspects of an evaluation to seven principles of quality. While the items and item descriptors for the evaluation quality tool may be further revised based on the feedback provided by members of the international education community during the evaluation review process, the results from this initial review process have already provided valuable insights into areas of strength and weakness. The process has also provided a first opportunity for the international education community to come together to discuss quality standards for USAID-funded evaluation with the Office of Education.

The Office of Education and its evaluation and implementing partners are likely to continue expanding the capacity of the Agency to consume, critique, and utilize high-quality reports with high methodological standards.[41] Thus, next steps may also include further developments of the evaluation quality tool to ensure its applicability to future innovations in the evaluation of USAID-funded interventions as well as periodic repetition of the evaluation quality assessment exercise.

Results from this evaluation quality assessment will also be used to determine which evaluation reports to include in the second phase of this study, which will synthesize findings and lessons learned about topics of interest to the Office of Education under each Strategy Goal.

[41] Andrew Green and Sam Hargadine, *Sectoral Synthesis of FY2015 Evaluation Findings: Bureau for Economic Growth, Education, and Environment* (USAID, December 2016).

ANNEX 1: STUDY STATEMENT OF WORK

Education Evaluation Syntheses – Goal 2

1. Activity Description

Building on recent efforts to synthesize what is being learned from evaluations that USAID commissions,[42] the Education Office in the Bureau for Economic Growth, Education, and Environment (E3/ED) is commissioning syntheses of evaluation findings related to the three Goals in the USAID Education Strategy. Products developed under this activity will address topics of interest to E3/ED and the Agency's education officers worldwide related to Goal 2 "Improved ability of tertiary and workforce development programs to generate workforce skills relevant to a country's development goals."

2. Existing Information Sources

E3/ED already has an inventory of recent education sector evaluations produced by the Bureau or by overseas Missions. Older evaluations, should the Bureau decide to examine a longer time period, can be accessed through the Agency's Development Experience Clearinghouse (DEC). Annual Performance Plan and Report (PPR) documents may be useful for determining whether evaluations reported as having been completed in previous years are all available in the DEC.

3. Activity Purpose, Audiences, and Intended Uses

Purpose

E3/ED intends that the analytic products that result from this activity will support evidence-based decision making by ensuring that findings from sets of evaluations on topics of interest to the Office are accessible to USAID staff. E3/ED's initial intent was to focus this activity on two topics related to Goal 2 of the Education Strategy: higher education and youth workforce development. Ensuing internal discussions led to an expansion of the scope to also include syntheses of evaluation findings on topics under Goals 1 and 3 under a common approach that could be applied across these three goals, and will be implemented across two mechanisms: the E3 Analytics and Evaluation Project and Reading and Access Evaluation. This activity will comprise two main phases. In phase 1, the quality of evaluation reports will be reviewed. In the phase 2, findings and lessons learned from a subset of evaluations that that met quality standards identified in the first phase will be extracted and synthesized. It is expected that up to 80 evaluation reports published between 2013 and 2016 across all three Goals will be reviewed under phase 1, with only a subset of those reports included in phase 2.

Audiences

The primary audience for the products to be developed under this activity are E3/ED and Mission staff as well as implementing and country partner organizations that plan and deliver education and workforce development programs and related support services.

[42] These efforts include the annual E3 Sectoral Synthesis of Evaluation Findings (https://www.usaid.gov/sites/default/files/documents/1865/E3_Sectoral_Synthesis_Report.pdf) and an evaluation synthesis from the Bureau for Food Security (BFS) that focuses on what has been learned from the Feed the Future initiative (https://agrilinks.org/sites/default/files/resource/files/Final%20KDAD%20Evaluation%20Sythesis.pdf).

Intended Uses

Two main reports are expected to be produced. The first report will be based on a standardized Evaluation Quality Protocol, and E3/ED will use its findings to determine topics on which it will develop additional guidance, products, and presentations to improve the quality of evidence generated for USAID-funded activities in the education sector. The second report will be based on a standardized Findings and Lessons Protocol, and E3/ED and Mission staff working the education sector may use the synthesized findings and lessons learned to inform future USAID education programming worldwide related to each of the three Education Strategy Goals. Work performed by the E3 Analytics and Evaluation Project under this activity should focus on Goal 2.

4. Synthesis Topics

E3/ED will confirm the topics for which findings and lessons learned will be extracted and synthesized. Tentative topics related to all three Education Strategy Goals are provided below:

Goal 1 – Early Grade Reading

Topic 1: Teacher training (pre-service and in-service)
Topic 2: Materials development, production, distribution, utilization
Topic 3: Parent/community engagement/support/education/mobilization
Topic 4: Systems/policy/government capacity strengthening

Goal 2 – Workforce Development

Topic 1: Training
Topic 2: Entrepreneurship
Topic 3: Private sector involvement
Topic 4: Systems/policy/government capacity strengthening
Topic 5: Youth engagement

Goal 2 – Higher Education

Topic 1: Training
Topic 2: Private sector involvement
Topic 3: Systems/policy/government capacity strengthening
Topic 4: Youth engagement

Goal 3 – Education in Conflict Settings

Topic 1: Training
Topic 2: Parent/community engagement/support/education/mobilization
Topic 3: Systems/policy/government capacity strengthening
Topic 4: Direct service delivery

5. Gender and Disability Considerations

Participation in the education system and educational outcomes vary considerably across countries, and can be substantially affected by gender and disability status. Therefore, it is expected that the syntheses prepared under this activity will report education-related findings by gender and disability status, when such information is available in the reviewed evaluations.

6. Activity Tasks

Evaluation Quality Protocol

In initial discussions about this activity, E3/ED requested that the synthesis team develop a preliminary framework for an approach to assessing the quality of evaluations to be examined under this activity. The framework the synthesis team prepared highlighted several core principles for consideration, including:

- Be consistent with USAID Evaluation Policy;
- Not be biased in favor of any particular evaluation design type, as it is expected that impact evaluations, performance evaluations, and qualitative evaluations will be reviewed; and
- Be amenable to a heterogeneous set of evaluation questions, ranging from the effectiveness of project/activity to the project/activity implementation and sustainability to the continued relevance of Agency assistance where circumstances may have shifted.

Pursuant to these recommendations, E3/ED requested that the synthesis team develop and pilot test an Evaluation Quality Protocol, which it will then pilot test in collaboration with E3/ED and incorporate feedback as appropriate. The protocol may also be shared with external audiences, such as the Comparative and International Education Society (CIES) annual conference, for additional feedback. This protocol should be used by a team of expert reviewers to identify which evaluation reports will have findings and lessons extracted and included in the syntheses.

The criteria for inclusion of reports in phase 2 will be developed by the synthesis team in collaboration with E3/ED. Possible criteria include the strength of the conceptual framing, openness, and transparency; robustness of methodology; cultural appropriateness; and the validity, reliability, and cogency of the evaluation. Data collected for phase 1 should also be analyzed and the main findings of this phase summarized in a 10- to 15-page report assessing the quality of the evaluations by Education Strategy Goal. Work performed under this activity should focus on the review of evaluations that have Goal 2 as their primary Education Strategy Goal. Results may also be disaggregated by geographic areas, pilots/scale-ups, whether conflict/crisis affected, and country income level.

Findings and Lessons Protocol

USAID experts in education sub-sectors will identify specific topics under each Education Strategy Goal about which findings and lessons learned will be synthesized (Section 4 provides a preliminary list of those topics). While the topics of the syntheses to be produced are expected to vary, E3/ED expects that the synthesis for each topic will be developed using a common outline, approach, and standards for documenting findings and lessons learned. Following initial discussions with E3/ED, the synthesis team will be expected to prepare a Findings and Lessons Protocol to extract findings (by gender and disability status when possible) and lessons learned (e.g., successes and challenges) about these topics from the evaluation reports. Existing E3 and BFS evaluation syntheses will be examined as potential templates, but final decision on the common outline, approach, and standards will be determined in collaboration with E3/ED. A Findings and Lessons Synthesis Report that addresses the agreed-upon topics under each Education Strategy Goal will be produced collaboratively by the two mechanisms. Work performed under this activity should focus on the extraction and synthesis of findings and lessons from evaluations that have Goal 2 as their primary Education Strategy Goal.

Team Selection and Training

Once E3/ED has selected the synthesis topics and approved the Evaluation Quality and the Findings and Lessons Protocols, the synthesis team will conduct training exercises for the reviewers. Different teams of reviewers might be used to apply each protocol. The selection and training of the reviewers for the Evaluation Quality Protocol should take into account the "Reviewing Evaluations for the Evaluation Synthesis Initiative" memorandum prepared by E3/ED, which suggests crowdsourcing the reviews in order to assess the quality of the evaluations while disseminating the evaluation quality criteria. The selection and training of reviewers for the Findings and Lessons Protocol should consider reviewers who are subject matter experts.

Implementation

Following E3/ED's approval of the Evaluation Quality Protocol and associated training, the synthesis team – in collaboration with E3/ED – will develop a systematic process for the review of the USAID evaluation reports. This process may include efforts to publicize the framework and quality criteria with key partners in the broader education and evaluation community. The implementation of the Evaluation Quality Protocol will result in the selection of the evaluations that will be subjected to the Findings and Lessons Protocol, based on criteria to be agreed with E3/ED as well as a summary report about the quality of the evaluations by Education Strategy Goal.

Following E3/ED's approval of the Findings and Lessons Protocol and associated training, the synthesis team – in collaboration with USAID staff in the topical fields on which the synthesis volumes will focus – will extract findings and other relevant data from topical sets of USAID evaluation reports produced between 2013 and 2016, and identify lessons for future programming, as relevant. This phase may also include the preparation of key findings and lessons summaries for each synthesis topic that would serve as a precursor to the preparation of a synthesis report and reviewed collaboratively by the synthesis team and USAID to highlight and prioritize the findings by topical area and identify any gaps in the summaries that may need to be addressed before a synthesis report is prepared.

Draft and Final Reports

The synthesis team will prepare drafts reports summarizing the main findings, conclusions, and recommendations from its implementation of the Evaluation Quality and Findings and Lessons Protocols. Based on USAID review and comments on such drafts, the synthesis team will prepare final versions for E3/ED's approval.

Dissemination Plan and Implementation

E3/ED will prepare dissemination plans for the reports produced under this activity, with inputs from the synthesis team as required. Thus, the schedule and budget for this activity should include time and resources for the synthesis team's involvement at the dissemination phase. The dissemination strategy should consider how the study products will be utilized by the identified audiences, and incorporate follow-up interviews as appropriate to determine and share actual instances of utilization.

7. Deliverables

A preliminary list of deliverables anticipated under this activity is provided below. The synthesis team, in consultation with E3/ED, will develop a Work Plan that will detail specific deliverables to be prepared under this activity with corresponding due dates. While products produced under this activity will focus on evaluations that have Goal 2 as their primary Education Strategy Goal, E3/ED might require that reports focusing on Goals 1, 2, and 3 be consolidated or summarized in one Evaluation Quality Report

and one Findings and Lessons Synthesis report to be prepared across the two implementing mechanisms.

1. Draft Activity Work Plan, including draft Evaluation Quality Protocol and draft Finding and Lessons Protocol
2. Final Activity Work Plan, including final Evaluation Quality Protocol and final Findings and Lessons Protocol
3. Draft Evaluation Quality Report, including draft dissemination plan
4. Final Evaluation Quality Report, including final dissemination plan and lessons learned about the evaluation quality review process and protocols
5. Draft Findings and Lessons Synthesis Report, including draft dissemination plan
6. Final Findings and Lessons Synthesis Report, including final dissemination plan and lessons learned about the findings and lessons review process and protocols

8. Team Composition

A research team led by Management Systems International (MSI) is expected to conduct this study across two mechanisms: the E3 Analytics and Evaluation Project, which is implemented by MSI in partnership with Development and Training Services and NORC at the University of Chicago; and the Reading and Access Evaluation project, which is implemented by NORC with MSI as a subcontractor. The review of evaluations and corresponding products related to Goal 2 will be conducted through the E3 Analytics and Evaluation Project while evaluations and corresponding products related to Goals 1 and 3 will be funded under the Reading and Access Evaluation project. Design, analysis, reporting, and dissemination efforts should be carried out across both mechanisms.

Separate Work Plans should be produced for the activities conducted under the E3 Analytics and Evaluation Project and the Reading and Access Evaluation project These Work Plans should propose a team and organizational approach to managing this activity, for E3/ED review and approval. It is recommended that the team include at a minimum an overall Team Leader, an Activity Coordinator, a designated Goal Lead for the Education Strategy Goal 2, as well as a sufficient number of mid- or senior-level Technical Advisors necessary to complete the tasks described in this SOW. It is expected that MSI will engage NORC at the University of Chicago to provide technical assistance and reviews of draft products at key points in this study.

9. USAID Participation

It is anticipated that E3/ED technical staff with expertise in the topics selected for examination under this activity will play an active role in developing the focus topics, reviewing study products, and developing lessons for future programming that will be incorporated into final syntheses volume(s). The exact nature of USAID staff participation will be further elaborated through discussions between E3/ED and the synthesis team, and may vary somewhat from topic to topic. In addition, through such discussions, E3/ED and the synthesis team will explore what roles implementing partners with which E3/ED collaborates may play in the topical areas to be covered.

10. Scheduling and Logistics

The tasks under this activity to be carried out by the E3 Analytics and Evaluation Project will be completed between approximately July 2016 and December 2017, with the timeline for subsequent dissemination tasks to be discussed with E3/ED. In its Work Plan, the Project team will propose a detailed schedule for implementation of the required tasks for this activity for E3/ED's approval.

11. Reporting Requirements

Reporting requirements will be finalized during discussions between E3/ED and the synthesis team concerning the synthesis topics, and will be incorporated into the final Work Plan.

12. Budget

The E3 Analytics and Evaluation Project team responding to this SOW will propose in its Work Plan an estimated budget to complete the tasks described in the Work Plan, for USAID's approval.

ANNEX 2: SELECTION OF EVALUATION REPORTS

The study team began the report selection process by providing the Office of Education with a list of potential evaluations for inclusion in this study. The study team generated the initial list from previous MSI work for the Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009-2012[43] (340 evaluations, 27 related to education), Sectoral Synthesis of 2013-2014 Evaluation Findings[44] (117 evaluations, 42 related to education), Sectoral Synthesis of FY2015 Evaluation Findings[45] (92 evaluations, 29 related to education), and Evaluation Utilization at USAID[46] (118 evaluations, 12 related to education). These studies identified USAID-funded evaluations through the Agency's Development Experience Clearinghouse (DEC) and Performance Plan and Reports. The study team subsequently identified additional evaluations through DEC searches as well as the Global Reading Network, YouthPower Learning, and the Education in Crisis and Conflict Network. The Office of Education also suggested evaluations and research studies that it deemed relevant for this study but which were not in the public domain.

The study only included evaluations of projects and activities that the Office of Education deemed relevant for the Education Strategy Goals. If there were multiple reports related to an evaluation of a single project or activity (e.g., baseline, midline, endline), the team only included the latest report. For evaluations of multi-country projects, the team included the available individual reports for each country. The Office of Education reviewed and vetted the final list of evaluations.

Once evaluations had been selected for inclusion, the study team worked with the Office of Education to identify the Education Strategy Goals associated with each evaluation. For evaluations of projects or activities covering multiple Education Strategy Goals, the team worked with the Office of Education to identify which was the primary Education Strategy Goal addressed by the evaluation. For the purposes of this study, the Office of Education instructed the study team to consider all evaluations that addressed school access as relevant to Goal 3, instead of only those in crisis or conflict environments. To identify crisis or conflict environments, the Office of Education provided the study team with a list of countries that were considered in crisis or conflict.

[43] See: http://pdf.usaid.gov/pdf_docs/pdacx771.pdf.
[44] See: https://www.usaid.gov/sites/default/files/documents/1865/E3_Sectoral_Synthesis_Report.pdf.
[45] See: http://pdf.usaid.gov/pdf_docs/PA00MP17.pdf.
[46] See: http://pdf.usaid.gov/pdf_docs/PA00KXVT.pdf.

ANNEX 3: EVALUATION QUALITY REVIEW STEPS

The evaluation quality review process included the following steps:

1. The Office of Education identified organizations to nominate staff to participate in the evaluation quality review.
2. The study team, in collaboration with the Office of Education, established the required minimum qualifications for reviewers to be identified by the partner organizations.
3. The study team, in collaboration with the Office of Education, issued organizational invitations to participate in the review on a volunteer basis.
4. Invited organizations nominated staff to participate in the review.
5. The study team, in collaboration with the Office of Education, confirmed if the proposed reviewers met the minimum qualifications.
6. The study team developed a web platform that included an interface for the tool with evaluation reports pre-loaded and prompts with item descriptors, which allowed reviewers to view and complete each review online, the team to monitor progress, and reviewers to submit feedback on the overall tool or on specific items.
7. The Office of Education and the study team co-presented a one-hour training webinar with the confirmed reviewers.
8. The study team uploaded to the web platform an orientation package including the webinar recording, a rater's guide for the tool, and the source of each item. The rater's guide included item descriptors with guidance on scoring all items in the tool (see Annex 4), which the team developed in collaboration with the Office of Education. Annex 7 contains the bibliographical information and link to the rater's guide shared with the reviewers;
9. The study team assigned evaluation reports to each reviewer and provided a timeline for completion of the reviews. Each reviewer was assigned two to three reports, and each report was reviewed by two reviewers from different organizations. To avoid the appearance of any potential conflicts of interest, reviewers were requested to inform the team if they were part of the evaluation or implementation team for their assigned report, or if they were employed by the company that did the evaluation or the implementation. In such cases, they were assigned to review a different evaluation. Reviewers completed the web application with their preliminary scoring, and the team monitored progress and followed up with reviewers as needed. Reviewers were not allowed to see each other's scoring during this part of the process.[47] The independent review process was prompted by the potential concern that the second reviewer could be less motivated to do a thorough review. When reviewers finalized and submitted their scoring, the web application locked their responses.
10. The study team hosted a full-day reviewers' meeting at MSI's offices in Arlington, Virginia. Twenty-six reviewers attended in person, and nine reviewers who were not closely located or who were otherwise unavailable to attend in person participated virtually. During this event, reviewers shared questions and discussed and provided feedback on applying the tool.
11. Following the reviewers' meeting, the team enabled the "harmonization" feature in the web platform that summarized discrepancies in responses of co-reviewers for the same evaluation. Once both co-reviewers had submitted their reviews, the system sent an automated email letting both co-reviewers know that the evaluation they assessed was ready for harmonization. For the harmonization, reviewers were shown their scorings side by side, with discrepancies highlighted in red, and they were asked to input their final responses in a consensus column.

[47] This differs from the GAO's performance audit of evaluation quality. In that audit, the first reviewer reviewed the evaluation, then notified the second reviewer that the evaluation report was available for a second review, with the second reviewer having access to the first reviewer's scoring. In this case, the second reviewer's main responsibility was to indicate whether he or she agreed with the first reviewer's scoring.

This consensus column was pre-populated with the responses that were the same across the preliminary scoring; however, reviewers could change responses for these items. Once reviewers discussed the evaluation and reached consensus on all items, they submitted their harmonized response, which was the final scoring for that evaluation. The study team monitored progress of these harmonizations and followed up with reviewers as needed.

Throughout this process, the study team substituted for the volunteer expert reviewers if certain reviewers were unable to complete their assignments. If reviewers could not harmonize their reviews, the team also reviewed the evaluation and served as an arbiter. MSI home and field office staff and consultants who met the reviewer qualifications for this study participated in this process.

ANNEX 4: TOOLS USED FOR THIS STUDY

Overview of the Tool Development

The study team developed and piloted with the USAID/E3 Office of Education an initial protocol that focused on a brief set of questions to capture "Evaluation Characteristics" and "Implementation Characteristics," as well as a detailed set of questions to capture "Methodological Quality" that it divided into two main domains—validity and reliability—and it applied to each finding in the evaluation report. The Office of Education then clarified that the framework used for this protocol should be expanded to include additional principles of quality as well as to increase its focus on education. Unlike other evidence rating systems such as the What Works Clearinghouse,[48] Clearinghouse of Labor Evaluation and Research,[49] and EVIRATER,[50] the principles of quality will be assessed for the overall evaluation instead of individual findings, similar to the GAO's performance audit on how Agencies Can Improve the Quality and Dissemination of Program Evaluations.[51]

The study team revised the protocol using the framework in the BE[2] guidance note on Assessing the Strength of Evidence in the Education Sector,[52] which focused on education research. This expanded the principles of quality that the protocol covered to include the evaluations' conceptual framing, openness and transparency, robustness of the methodology, cultural appropriateness of the tools and analysis, and cogency of the report, in addition to validity and reliability. This framework also elicited overall ratings for the evaluation instead of for individual findings. The topics addressed under each of the principles of quality are summarized below.

- **Conceptual Framing**
  - Theory of change
- **Openness and Transparency**
  - Design and methods
  - Data analysis
- **Robustness of methodology**
  - Appropriateness of design
  - Rigorous application
- **Cultural appropriateness/sensitivity**
  - Culturally relevant tools
  - Culturally sensitive analysis
- **Validity**
  - Measurement, internal, external, and ecological validity
- **Reliability**
  - Consistent measurement
  - Consistent results from repeated processing and analysis
- **Cogency**
  - Logical argumentative thread throughout the entire paper
  - Conclusions based on results

[48] See: https://ies.ed.gov/ncee/wwc/.
[49] See: https://clear.dol.gov/.
[50] See: http://abtassociates.com/Noteworthy/2015/EVIRATER-Rating-the-Strength-of-Evidence-in-Evalua.aspx.
[51] See: http://www.gao.gov/assets/690/683157.pdf.
[52] See: https://www.usaid.gov/sites/default/files/documents/1865/BE2_Guidance_Note_ASE.pdf.

The protocol incorporated all BE^2 recommended associated principles, which are phrased as questions such as "Does the study acknowledge existing research?" or "Does the study demonstrate why the chosen design and method are good ways to explore the research question?" and are rated as "low," "mid," or "high" and then used to produce a final rating for the evaluation as "low", "mid", "high," or "very high". The team then added questions based on USAID guidance regarding evaluation reports,[53] the evaluation report quality checklist used in the E3 Sectoral Synthesis of Evaluation Findings,[54] and the Critical Appraisal Skills Programme Qualitative Checklist.[55] The team complemented these with questions based on guidance from the What Works Clearinghouse's Procedures and Standards Handbook,[56] Running Randomized Evaluations: A Practical Guide,[57] and the Early Grade Reading Assessment Toolkit: Second Edition.[58] The team mapped all added questions to the appropriate quality principle, to prepare reviewers before they rated the BE^2 associated principles under each quality principle, thus making this iteration a BE^2-*plus* protocol.

At the suggestion of the Office of Education, the team also explored developing modules to apply to specific issues, such as the evaluations' compliance to Goal 1, or to apply to specific measurement instruments such as an early grade reading assessment checklist. The team also expanded the brief set of questions that captured "Evaluation Characteristics" and "Implementation Characteristics" and, as suggested by the Office of Education, moved these questions to a separate tool to be filled out by non-experts (which it also piloted with the Office of Education).

The review process thus included two tools: a non-expert tool to extract basic information about each evaluation, and a tool for technical experts to assess the quality of each evaluation.

Based on the results of final piloting with the Office of Education, the team and the Office of Education made changes to the evaluation quality assessment tool, which included: dropping the modules for specific issues and measurement tools; reorganizing the questions' so they do not lead to the BE^2 associate principles within each principle of quality but rather integrate these as regular questions; rephrasing and dropping questions; capturing expert judgments and justification about whether the evaluation adequately addressed each principle of quality; and dropping an overall rating for the evaluation. The Office of Education and the team then co-presented the evaluation quality assessment tool at a workshop during the CIES 2017 annual conference. During this workshop, attendees from USAID implementing and evaluation partner organizations, as well as from universities, piloted the tool and provided feedback. After CIES, the team and the Office of Education incorporated this feedback into the tool, including shortening it to 40 questions (4 to 8 questions per principle of quality) plus the expert judgment and accompanying justification about whether the evaluation adequately addressed each principle. The team checked the questions in the revised tool for alignment with those in the GAO report, Agencies Can Improve the Quality and Dissemination of Program Evaluations.[59]

[53] This guidance includes USAID Evaluation Policy, USAID Scientific Research Policy, and relevant Automated Directives System (ADS) sections for evaluation including ADS 201maa "USAID's Criteria to Ensure the Quality of the Evaluation Report," ADS 201mah "USAID Evaluation Report Requirements," and ADS 201sae "USAID Data Quality Assessment Checklist and Recommended Procedures."
[54] See: http://pdf.usaid.gov/pdf_docs/pa00mp17.pdf.
[55] See: http://media.wix.com/ugd/dded87_25658615020e427da194a325e7773d42.pdf.
[56] See: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf.
[57] See: http://runningres.com/.
[58] See: http://pdf.usaid.gov/pdf_docs/pa00m4tn.pdf.
[59] See: http://www.gao.gov/assets/690/683157.pdf.

Tool for Background Information about Evaluations

Questions	Score	Description
[1] Is the evaluated intervention a pilot or a large-scale implementation?	pilot/full intervention/not applicable	The scope of implementation of the project that is being evaluated. A pilot is a smaller scale project that enables decision makers to "try out" an activity before deciding whether and how to roll out the activity at a larger scale. Large scale implementation generally touches a greater number of beneficiaries and does not assume a "trial" of activities. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[2] Geographic coverage of the evaluation	multi-country/national/sub-national/district	Does the project focus on several countries (multi-country), at the central level of a country, at its subnational level (e.g. state, province) or its the district level?
[3] Type of evaluation	impact, experimental/ impact, quasi-experimental/ impact, non-experimental/ performance, quantitative/ performance, qualitative/ study	Please choose the appropriate evaluation type from the dropdown menu. Impact evaluations: Measure the change in a development outcome that is attributable to a defined intervention; impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change. Impact evaluations in which comparisons are made between beneficiaries that are randomly assigned to either a treatment or a control group are experimental evaluations. When assignments are not random, the evaluations are quasi-experimental. Performance evaluations: Encompass a broad range of evaluation methods. They often incorporate before-after comparisons, but generally lack a rigorously defined counterfactual. Performance evaluations may address descriptive, normative, and/or cause-and-effect questions: what a particular project or program has achieved (at any point during or after implementation); how it is being implemented; how it is perceived and valued; whether expected results are occurring; and other questions that are pertinent to design, management, and operational decision-making. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[4] Stage of evaluation	baseline/mid-term or midline/final or endline/not applicable	The stage of the project where the evaluation is taking place. Evaluations typically occur prior to or at an early stage of implementation (baseline); during the course of the project, between the beginning and end (mid-term/midline); or at the completion stage of a project (final/endline). "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[5] Staffing of evaluation team (Mark all that apply)	international evaluation specialists/ local evaluation specialists/ representative of the implementation team/ USAID staff/	The type of staffing used to complete the evaluation. Multiple selections are possible. An international evaluation specialist or expert does not come from the country for which the evaluation is being conducted, whereas a local evaluation specialist does. At times, participatory evaluations are conducted, with a USAID staff member, an international stakeholder, and/or other local stakeholders joins the evaluation team. To be a part of the evaluation team, they

Questions	Score	Description
	other international stakeholders/ other local stakeholders/ not reported	must go beyond simply contributing data. "Not reported" score should be given if the report doesn't report the evaluation team's composition.
[6] Are existing project's monitoring data included in the evaluation data analysis?	yes/no/not applicable	Performance monitoring is the ongoing and systematic collection of performance indicator data and other quantitative or qualitative information to reveal whether implementation is on track and whether expected results are being achieved. Performance monitoring includes monitoring of outputs and project and strategic outcomes. Monitoring data may be taken from the project's quarterly, annual or final reports. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[7] Are fidelity of implementation data from the project M&E system included in the evaluation data analysis?	yes, data from the assessment/no/not applicable	"Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[8] Does the report include information about audience for the report?	yes/no	The report should have an audience for whom the evaluators wrote the report. Examples include, the USAID mission, USAID/Washington, Government, other donor groups, implementing partners, etc. This information is typically included in the introductory/background section.
[9] Does the report explain how findings will be used?	yes, in detail/yes, but little detail/no	A management purpose of the evaluation should be explicit in regards to the decisions and actions the evaluation is intended to inform. An evaluation can also have more than one management purpose.
[10] Does the report include a description of the intervention?	yes, in detail/yes, but little detail/no/not applicable	The project description plays a critical role in enabling the reader to understand the context of the evaluation, and involves several characteristics such as the title, dates, funding organization, budget, implementing organization, location/map, and target group. All of these characteristics play an important role and virtually all should be present to receive credit for this item in order to take a holistic view of whether the project is sufficiently well-described. If most characteristics are present, as well as information on how the project was carried out, then check "yes, in detail". If a number of characteristics are missing or weak, or if information on how the project was carried out isn't offered, then check "yes, but little detail." If not information is provided, check "no". "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[11] Does the report include information about intervention dosage at the beneficiary level?	yes, for all beneficiaries/yes, for some beneficiaries/no/not applicable	Intervention dosage may refer to the reach of project activities, including the number of beneficiaries touched, as well as key output information, such as number of teachers trained for example. This does not relate to outcome indicators. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.

Questions	Score	Description
[12] Does the report include project's logic model/results framework/theory of change?	yes/no/not applicable	The "theory of change" describes, via narrative and/or graphic depiction of the intended results and causal logic, how anticipated results will be achieved. You may see this described as the development hypotheses and assumptions underlying the project or program. We expect that a clear explanation of the theory of change/development hypotheses will be presented in the evaluation report before the evaluation's finding are presented. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[13] Does the report include information about project targets?	yes/no/not applicable	Performance targets are a key aspect of project monitoring. They are defined as the specific, planned level of a result to be achieved within an explicit timeframe with a given level of resources, and are expressed quantitatively. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[14] Evaluation respondents (Mark all that apply)	pre-school age children/ primary grade students/ out-of-school children age 6 through 14/ secondary grade students/ vocational school or tertiary level students/ non-formal education or alternative education learners/ out of school youth (age 15-24)/ parents/ early childhood educators or master trainers/ upper primary or secondary school educators or master trainers/ tertiary or vocational instructors or master trainers/ government officials or administrators/ entrepreneurs	Who were the people or institutions that the evaluators contacted to gather data? Check all that apply.
[15] Were learning assessments included as part of the evaluation? (Mark all that apply)	none/ early childhood assessment/ early grade reading/ early grade math/ vocational skills/ soft skills or social-emotional skills	The type of learning assessment that was used to evaluate the intervention. Early childhood assessments may include tests or ratings. For early grade reading, the Early Grade Reading Assessment (EGRA) is most commonly used, and for math it is the Early Grade Math Assessment (EGMA). ASER and UWEZO have also developed assessments that can be used for literacy or math. Vocational skills for youth, such as computer skills, or a technical trade, may also be assessed in an evaluation. Other important skills that may be assessed include soft/social-emotional skills.
[16] Were other assessments included as part of	none/ institutional capacity assessment (ICA)/	Check all that apply. In addition to learning assessments, other assessments may be conducted, such as an institutional or organizational capacity assessment. An

Questions	Score	Description
the evaluation? (Mark all that apply)	organizational capacity assessment (OCA)/ assessment of school management or leadership/ assessment of teacher knowledge or practice/ assessment of teacher wellbeing or motivation/ assessment of learning environment, including safe learning environment or GBV/ assessment of learners' wellbeing/ assessment of community or parents or caregivers	assessment of school management/leadership may assess for example master trainers/principals/head teachers or hiring/payment/training of teachers. An assessment of teacher knowledge and practice can often be in the form led through a classroom observation of teaching techniques. An assessment of teacher wellbeing may consider teacher's training levels, resource availability and morale. An assessment of learning environment or safe learning environment may consider the resources available in a classroom, the school structure, the proximity to safe drinking water, but also the punitive methods in class, distance to home, gender separated bathrooms etc. An assessment of learner wellbeing typically evaluates a learner's psycho-social state. Finally, an assessment of community/parents/caregivers may aim to determine their engagement with the school life and involvement in the students' education.
[17] Are the evaluation results disaggregated by population subgroups? (Mark all that apply)	none/ gender/ socio-economic status/ ethnic or linguistic group/ disability status/ grade level/ religion	Because data may hide disparities when looking at the aggregate, it is important to examine how data differs among different groups. Indicate if data is provided that is disaggregated by gender, socio-economic, ethnic/linguistic, disability, grade level, and/or religious groups. Check all that apply.
[18] Do evaluation questions include a deliberate and explicit exploration of gender, disability, ICT in education, innovative financing, or scale and sustainability topics? (Mark all that apply)	no/ yes, gender/ yes, disability/ yes, ICT in education/ yes, innovative finance/ yes, scale and sustainability	Other topical areas of interest include gender, disability, Information and Communication Technology (ICT) in education, innovative financing, or scaling-up and sustainability of projects. Check all that apply.

Evaluation Quality Tool: Working Version

Questions	Score	Description
Conceptual Framing		
[1] Are the research/evaluation questions included in the report?	yes/no	All research/evaluation questions must be phrased as questions; it is not enough that they be inferable from the stated objectives of the study. Questions must be clearly stated and be answerable through the reported research methods.
[2] Does the report include research/evaluation hypotheses?	yes/no/not applicable	Research/evaluation hypotheses must be explicitly described; it is not enough that they be inferable from the stated objectives of the study. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[3] Are the research/evaluation questions appropriate for the intervention's conceptual framework (logframe/theory of change/results framework)?	yes/partial/no /not applicable	All research/evaluation questions should be based on the intervention's conceptual framework. "Partial" score could be given when some, but not all, listed evaluation questions correspond to the intervention's conceptual framework. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[4] Does the report acknowledge/draw upon existing country-specific research?	yes/partial/no	Studies should build on existing research, both local and funded by international donors. The report should specify how questions, methodology, tools and analysis plans are informed by prior research. "Partial" score could be given when only some of the questions are informed by existing knowledge.
[5] Does the report explain the local context in sufficient detail?	yes/partial/no	The local context should be explained in enough detail for a general audience to be able to appreciate the relevance of the intervention being evaluated. "Partial" score could be given when some, but not all, elements of the intervention have corresponding contextual information.
[6] Conceptual framing: Conclusion	adequate/not adequate	**Adequate:** Overall, this evaluation demonstrates adherence to principles of conceptual framing **Not Adequate:** This evaluation contains major deficiencies in demonstrating adherence to principles of conceptual framing or provides insufficient information for determining this
[7] Conceptual framing: Notes/Justification		For instance: *"The authors acknowledge existing research and make clear how their analyses sit within the context of existing work. They provide a theoretical framework in the report, where they outline their major assumptions. The study also poses specific research questions."*
Openness and Transparency		
[8] Is the report open about study limitations/weaknesses due to the methodology, sampling, data collection, etc.?	yes/partial/no	It is common for evaluators to encounter expected or unexpected interferences with study design or the implementation of the study. Evaluators are obligated to include these "study limitations" and a description of the impact they may have had on the evaluation. Clarity around study limitations is particularly important only if they directly impact the evaluator's ability to credibly and effectively answer an evaluation question or impact generalizability of the findings (i.e., if data collection was successful but more expensive or inconvenient than anticipated, it is not a limitation). "Partial" score could be given if the report mentions common limitations, such as potential issues with the generalizability of the findings, without discussing them in detail.

Questions	Score	Description
[9] Is the report open about study limitations due to issues with the implementation of the intervention being evaluated?	yes/partial/no /not applicable	Interventions frequently evolve in a way that may compromise the integrity of the evaluation design. For instance, a new component of the intervention may be introduced midway through the implementation. Another example might be poor records of the implementation itself making it impossible for the evaluators to establish to what the observed effects might be attributed. Any such limitations of the intervention itself (not the evaluation) should be reported and their implications for the evaluator's ability to credibly answer the evaluation question discussed. "Partial" score could be given if the report mentions common limitations, such as potential issues with program implementation, without discussing them in detail. "Not applicable" score should be given to research studies that do not evaluate a specific intervention.
[10] Does the report include alternative interpretations of the findings?	yes/no	The evaluation report should balance the presentation of the findings with the inclusion of alternative explanations for the findings. Some reports may even include alternative causes. If so, it is important that the evaluators report such information, and, if these yield inconsistencies with other results, that these are also pointed out.
[11] Does the report present the complete analysis of data relevant for study questions?	yes/partial/no /not applicable	The evaluation report should make it clear which data collection and data analysis methods were used to analyze data to answer specific evaluation questions, and all results should be included in the report (e.g. summary of focus groups, zero scores, breakdown by gender). Detailed data analysis results may be available within the body of the report or may be found in an annex. "Partial" score could be given if the report includes detailed data analysis results for some, but not all evaluation questions. "Not applicable" score could be given if no evaluation questions are provided in the report.
[12] Is the report open about potential biases due to the study team composition?	yes/partial/no	USAID encourages study teams to include at least one evaluation specialist, host country team members, and a team leader who is external to USAID. USAID also requires that evaluation team members certify their independence by signing statements indicating that they have no conflict of interest or fiduciary involvement with the project or program they will evaluate. It is expected that an evaluation will indicate that such forms, or their equivalent, are on file and available or are provided in an evaluation annex. "Partial" score could be given if some, but not all, these recommendations are followed.
[13] Openness and transparency: Conclusion	adequate/not adequate	**Adequate:** Overall, this evaluation demonstrates adherence to principles of openness/transparency **Not Adequate:** This evaluation contains major deficiencies in demonstrating adherence to principles of openness/transparency or provides insufficient information for determining this
[14] Openness and transparency: Notes/Justification		For instance: *"The authors are transparent about the design and methods that have been employed in the evaluation as well as the data (and resulting sample) that have been gathered and analyzed. This allows for the study to be repeated and corroborated."*

Questions	Score	Description
Robustness of Methodology		
[15] Is the methodology explained in sufficient detail?	yes/partial/no	USAID requires that an evaluation report identify the study design, data collection methods and data analysis techniques used. It is common to include the methodology description in the body of the report under a methodology section with a longer and more detailed methods annex. The description of methods must indicate how respondents were selected, what types of interviews were conducted, with whom they were conducted (e.g., key informant interviews, individual interviews with beneficiaries, group interviews), as well as detailed information on the kinds of analyses that were conducted (e.g., correlations, regressions, content analysis, pattern analysis). "Partial" score could be given if some, but not all elements mentioned (design, data collection methods and data analysis techniques) were described in sufficient detail.
[16] Is the methodology appropriate for answering posed study questions?	yes/partial/no	USAID recognizes that different designs are more or less appropriate to answering different research questions, and that the selection of method (or methods) for a particular evaluation also balances cost, feasibility, and the level of rigor needed to inform specific decisions. Thus, USAID Evaluation Policy is clear that no single evaluation design will be privileged over others: Observational, quasi-experimental, and experimental designs all may yield valuable findings, but are appropriate for answering different types of questions. Assessing the appropriateness of the chosen methodology is complicated by the fact that most evaluations include a variety of questions that most frequently require a mixed-method approach, and the assessment of the methodology must include the review of the evaluation design vis-a-vis stated study questions. "Partial" score could be given if the methodology proposed is appropriate for some, but not all posed questions.
[17] Does the counterfactual meet standards of rigor?	yes/no/not applicable	Measuring what would have happened in the absence of an intervention is a requirement for establishing a causal relationship. A counterfactual can be created in a number of ways, from simply using respondents from a geographically close unit as comparison group to using statistical analysis to compensate for the potential selection biases of non-randomization to randomly assigning subjects to treatment(s) and control groups. "Not applicable" score should be given if the evaluation/research is not an Impact Evaluation
[18] Is data triangulation described as part of methodology?	yes/partial/no/not applicable	Typically, stronger bodies of evidence are likely to emerge if similar findings are obtained from different types of data (e.g., tests, interviews, observations) and respondent types (e.g., students, parents, teachers). It is important that contradictory data be taken into account when discussing the findings. "Partial" score could be given if data from different sources are presented but the findings don't connect them into a coherent narrative. "Not applicable" score should be given if the evaluation/research is not a Performance Evaluation
[19] Does the report mention steps to mitigate common threats to the integrity of the evaluation	yes/partial/no/not applicable	USAID Evaluation Policy requires that evaluation reports address methodologically common limitations, such as when there is a disjunction between the treatment that is assigned and the treatment that is received (non-compliance). "Partial" score

Questions	Score	Description
(such as non-equivalence at baseline, non-compliance, spillover, systematic attrition) or common biases (confounding bias, selection bias, experimenter bias, etc.)?		could be given if some, but not all threats or biases identified are discussed. "Not applicable" score could be given if no threats or biases were identified
[20] Are sampling approach and sample size calculations presented in sufficient detail (to include, at a minimum, type of analysis, MDES, alpha and beta)?	yes/partial/no/not applicable	Details of power calculation should be included in either the main body of the report or in an annex. This should include the parameters used in the power function that relates power (beta) to its determinants: (1) level of significance (alpha), (2) minimum detectable effect size (MDES) or minimum detectable impact (MDI), (3) and the sample size. "Partial" score could be given if the description of the sample size calculations presents only some of the parameters used. "Not applicable" score could be given if the evaluation/research used only qualitative research methods
[21] Is the sampling approach described in sufficient detail? (at a minimum, a rationale for the sample size and method of sample selection) and is it appropriate for the study objectives?	yes/partial/no/not applicable	Researchers/evaluators should provide a description of the sampling frame and potential issues with it, if any. This should include an explanation of how the participants were selected, whether these participants were the most appropriate to provide access to the type of knowledge sought by the study, whether there was a point at which incoming data produced little or no new information (saturation) as well as any discussions around recruitment, such as why some people might have chosen not to take part in the study. "Partial" score should be given if only some of these elements were discussed. "Not applicable" score should be given if this study did not use qualitative research methods.
[22] Robustness of methodology: Conclusion	adequate/not adequate	**Adequate:** Overall, this evaluation demonstrates adherence to principles of appropriateness/rigor of chosen methodology **Not Adequate:** This evaluation contains major issues with the appropriateness of the chosen methodology, major deficiencies in the rigor with which it was applied or provides insufficient information for determining this
[23] Robustness of methodology: Notes/Justification		For instance: *"The study aims to identify and examine specific effects of receiving grants alone compared to receiving grants as well as training on student learning outcomes. The study clearly aims to establish a causal linkage between grants versus grants/training on student outcomes. The experimental design was, therefore, most appropriate to answer the research question. The study demonstrates rigorous application of the experimental technique within The Gambian setting. The authors clearly describe the interventions and adopt all the rigors of a well-applied randomization."*

Questions	Score	Description
Cultural Appropriateness		
[24] Does the report include documentation of local ethics review and/or US-based IRB approval/exemption status?	yes/no/not applicable	As outlined in the USAID Scientific Research Policy, USAID-funded research/evaluations must conform to legal and other requirements governing research with human subjects in the country where it is conducted. To satisfy USAID CFR requirement of ethics review, a study must be reviewed and approved or deemed exempt by a US-based IRB. USAID accepts legitimate foreign procedural systems in lieu of the U.S.-based IRB review only when they are determined to provide protection "at least equivalent" to the Common Rule. "Not applicable" score should be given if request for approval or exemption status is not needed, such as research that does not involve human subjects (e.g. secondary data analysis).
[25] Does the report list steps taken to ensure that study questions and methodology are informed by local stakeholders, are culturally relevant and appropriate?	yes/no	The evaluation questions and methodology should be informed by relevant local stakeholders. This could be done during in-country design workshops as well as through meeting with the ministry or other relevant stakeholders.
[26] Does the report list steps to ensure that data collection tools were developed/adapted with participation of relevant local stakeholders and are culturally appropriate?	yes/partial/no	The report should describe whether tools have been developed to suit the local context, such as whether the tool was developed by international experts and then merely translated into a local language or whether local knowledge has been used effectively in the adaptation of the tool to reflect resources relevant to the context, such as including support from host country experts. "Partial" score could be given if some, but not all tools suit the local context.
[27] Does the report list steps taken to validate findings/conclusions/recommendations with local stakeholders?	yes/no	Findings, conclusions and recommendations must be communicated to the appropriate audiences in a culturally and contextually suitable way prior to finalization of the report, in order to validate accuracy of conclusions and help inform recommendations. Steps to validate these with local stakeholders may include in-country presentations and workshops.
[28] Is the study informed by locally relevant stratifiers, such as political, social, ethnic, religious, geographical or sex/gender phenomena (including methodology, data collection and data analysis)?	yes/partial/no	The extent to which a study takes into account locally relevant stratifiers has considerable bearing on the study's design, its analytical strategy and the interpretation of its findings. Being informed by locally relevant stratifiers might include making cross-cultural or cross-linguistic comparisons part of the analytical strategy or ensuring that knowledge of the local context is used in the interpretation of differential effects between groups. "Partial" score should be assigned when the study is purposeful with considering variable impacts on gender but not any other stratifiers.
[29] Cultural appropriateness: Conclusion	adequate/not adequate	**Adequate:** Overall, this evaluation demonstrates adherence to principles of cultural appropriateness. **Not Adequate:** This evaluation contains major deficiencies in demonstrating adherence to principles of cultural appropriateness or provides insufficient information for determining this.

Questions	Score	Description
[30] Cultural appropriateness: Notes/Justification		For instance: *"The evaluation describes systematic processes used to check for the cultural relevance of measurement items (for example, in the absence of lists of age-specific words for Bangla-speaking children, a list was created of words that fit two criteria: they should be known to grade 1 or 2 children but unknown to preschoolers, and they should be used in the storybooks). Thus, the instrument used is culturally sensitive. The analysis is also culturally sensitive, as it discusses the factors that undermine or promote educational outcomes within the Bangladeshi context. The study discusses the use of two supply-and-demand side interventions – a school-only grant and a school grant plus an education allowance – which the authors discuss in relevance to the context, where grants are used to provide key inputs to schools while the education allowance provides a conditional monetary incentive for out-of-school children to attend school."*
Validity		
[31] Do indicators used in the evaluation serve as appropriate proxies for the construct or phenomenon being investigated?	yes/partial/no	In order to assess the validity of the measurement, it is important to consider whether or not the chosen indicators adequately capture the concepts being measured or whether there are other dimensions central to the concepts that are being ignored, such as a labor market condition index that ignores underemployment. "Partial" scores could be given if some, but not all key indicators, adequately captured the concepts being measured.
[32] Were the assessments conducted in such a way such that the results are generalizable to the population of students reached through the activity?	yes/partial/no/ not applicable	A number of characteristics of the survey design, such as timing of the assessment and absence of sampling weights, may affect the interpretation and/or calculation of population estimates. The evaluator/research may provide information about the timing of the assessment (e.g., pre-test and post-test being conducted at comparable time points in a cross-sectional design) or construction and use of sampling weights in the analysis (when different observations in a random selection process may have different probabilities of selection). "Partial" score could be given if the report mentions that the interpretation and/or calculation of some but not all population estimates took into account relevant survey design characteristics. "Not applicable" score should be given in case this is a qualitative study.
[33] Does the report allude to whether the study findings may have been biased by the activity of doing the study itself?	yes/no	Evaluators/researchers might discuss in the report whether findings could have been influenced by the process of research itself (ecological validity) or whether participants may have changed their behavior in response to their perception of the evaluators' objective (response bias), such as when the treatment group works harder than normal in response to being part of an evaluation (Hawthorne effects). Note that the tendency of participants to give an answer to a question that is in line with social norms even if this does not accurately reflect their experience (social desirability bias) is not relevant for this question.
[34] Does the report address the external validity of findings?	yes/partial/no	Findings are externally valid when they are valid in contexts other than those the evaluation was conducted in. Thus, researchers/evaluators may discuss the local conditions that

Questions	Score	Description
		would make it replicable in a different context. "Partial" score could be given if the external validity of some, but not all key findings, are discussed in the report.
[35] Were all data collection tools piloted with representatives of target populations prior to beginning of the data collection?	yes/partial/no	Researchers/evaluators should describe if respondents used to pilot the data collection tools were similar to the target population of the full study. "Partial" score could be given if the report mentions that piloting was done but not with who.
[36] Are confidence intervals reported around point estimates?	yes/no/not applicable	USAID recommends that the margin of error be reported along with the findings from statistical samples. "Not applicable" score should be given if the study does not use inferential statistical methods.
[37] Are reported relationships tested for statistical significance and *p*-value reported?	yes/no/not applicable	Evaluators often use statistical tests such as t-test and F-tests to determine whether an estimated coefficient or effect is statistically different from a specified value (usually zero) or whether two numbers are significantly different from each other. The results of these significance tests of probability value should be provided in the report. The *p*-values may also be adjusted to account for the fact that several different hypotheses are being tested in the study (e.g. Bonferroni correction). "Not applicable" should be given if the study does not use inferential statistical methods.
[38] Is treatment effect presented in terms of effect size?	yes/no/not applicable	Researchers/evaluators often record the study findings in the units of the outcome variable. To improve the comparability of effect size estimates across outcome variables and across studies, effect sizes in terms of standard deviations should also be provided, taking into consideration the study design. "Not applicable" should be given if the study did not conduct statistical hypothesis testing (as in the case of qualitative studies).
[39] Validity: Conclusion	adequate/not adequate	**Adequate:** Overall, this evaluation demonstrates adherence to principles of validity. **Not Adequate:** This evaluation contains major deficiencies in establishing the measurement, internal, external or ecological validity or provides insufficient information for determining this.
[40] Validity: Notes/Justification		For instance: *"The authors describe steps they took to address the validity of the study. For example, items included in the test had to relate directly to what grade 5 children would be expected to know at the start and end of the school year and statistical analyses were conducted to assess the internal consistency of questions in order to refine and adjust the assessment tools (measurement validity). In assessing learning progress of pupils in grade 5, the study included initial test scores into the estimation and controlled for background factors that may generate biases (internal validity). The study is based on longitudinal data collected from 5 provinces out of 58 in Vietnam, the generalizability of the findings is somewhat questionable (external validity), and there is no discussion of whether the findings could have been influenced by the process of research itself (ecological validity). While it could be improved, overall this study meets basic standards of scientific validity."*

Questions	Score	Description
Reliability		
[41] Does the report list steps taken to ensure that data were collected with a high degree of reliability?	yes/partial/no	USAID recommends that data collection methods be documented in writing to ensure that the same procedures are followed each time. The report may describe the use of data quality assurance checks such as accompaniments, back-checks and scrutiny, and these may have been conducted through spot-checking or for all questions in the data collection form. In case of paper-and-pencil data collection, double data entry report and/or double manual verification may also be mentioned in the report. Steps used in qualitative studies may include audio recording, videotaping and transcribing interviews. "Partial" score could be given if steps to ensure the reliability of some, but not all data collected, are described.
[42] Does the report provide statistics on inter-rater reliability of assessors during field data collection?	yes/no/not applicable	Inter-rater reliability statistics (like raw agreement and kappa) are measurements of the consistency between assessors. The USAID/E3 Office of Education recommends that in addition to an assessor evaluation process during training, that researchers/evaluators have two or more assessors in a sample-base collect data from the same respondent at the same time to compute the inter-rater reliability statistics for the field data collection. "Not applicable" score should be used for qualitative studies.
[43] Are data collection tools included in the report?	yes/partial/no	USAID recommends that all data collection tools used, such as questionnaires, checklists, survey instruments, and discussion guides, be included in the report's appendix. "Partial" score should be given if some, but not all, data collection tools are provided in the appendix.
[44] Does the report adequately address missing data/non-response?	yes/partial/no	Researchers/evaluators should report the target number of respondents, the number of respondents reached, and the number of respondents who were included in the data analysis. This includes non-response in qualitative studies. For quantitative evaluations, the report may also mention using post-stratification to adjust weights for non-response. "Partial" score could be given if information about valid responses is provided to some, but not all data used in the findings.
[45] Reliability: Conclusion	adequate/not adequate	**Adequate:** Overall, this evaluation demonstrates adherence to principles of reliability. **Not Adequate:** This evaluation contains major deficiencies in establishing the reliability of the measurement or provides insufficient information for determining this.
[46] Reliability: Notes/Justification		For instance: *"This study used multiple researchers to undertake school observations and interviews; the researchers checked their own conclusions with each other and then cross-checked them against the wider analytical team to analyze between schools. The team ensured that different types of data were collected – observations, interviews and document analysis – to triangulate findings and take into account the variety of possible contexts. The authors also provide a good example of how to enhance the reliability of qualitative analysis: interviews were videotaped and transcribed."*
Cogency		
[47] Are all of the study questions, including sub-questions, answered?	yes/no/not applicable	The purpose of an evaluation report is to provide the evaluators' findings and recommendations on each and every evaluation question. Accordingly, USAID expects that the answers to all

Questions	Score	Description
		evaluation questions, including any sub-questions, will be provided in the report. "Not applicable" score could be given if no evaluation questions are provided in the report.
[48] Does the Executive Summary include answers to all of the study questions?	yes/no	The executive summary must provide an accurate representation of the main elements of the evaluation report without adding any new material information or contradicting the evaluation report in any way. As such, it is recommended that all evaluation questions/issues, including any sub-questions/issues, will be provided in the Executive Summary.
[49] Is the report written in a language accessible to the audiences for whom the report indicates it is written?	yes/no	Reports should be written in a language understandable to non-researchers and non-American audiences. Excessive use of research terminology is also undesirable; the report should favor terminology that its intended audience is expected to be familiar with.
[50] Are recommendations action-oriented, practical and specific?	yes/partial/no/ not applicable	USAID requires evaluation teams to include information about evaluation audiences and a utilization plan. "Partial" score could be given when some, but not all, recommendations listed identify the specific actions recommended to the specific party. "Not applicable" score could be given if no recommendations are presented in the evaluation report, such as in baselines reports (if at this phase, recommendations for the intervention would not be appropriate) or in research studies.
[51] Is there a transparent connection between the study questions, findings from the data and the conclusions and recommendations, and is the report structured to present findings clearly and objectively?	yes/partial/no	USAID requires that evaluation findings be based on reliable quantitative and/or qualitative data, and that conclusions and recommendations should be based on these findings. USAID also encourages evaluators to present a clear progression from Study questions to Findings to Conclusions to Recommendations (if any) in their reports, such that none of a report's conclusions and recommendations appear to lack grounding. "Partial" score could be given if some supporting data is provided for some, but not all findings.
[52] Are visuals in the report appropriate for helping non-technical audiences easily understand the study findings?	yes/partial/no	Visuals must be used to facilitate understanding of the findings by general audiences. Visuals should be standalone, such that they are interpretable without the audience needing to read extra text. "Partial score" could be given if the report uses visuals to an insufficient extent.
[53] Cogency: Conclusion	adequate/not adequate	**Adequate:** Overall, this evaluation demonstrates adherence to principles of cogency. **Not Adequate:** This evaluation contains major deficiencies in demonstrating adherence to principles of cogency or provides insufficient information for determining this.
[54] Cogency: Notes/Justification		For instance: *"The evaluation contains a clear, logical argumentative thread that runs through the entire report. This links the conceptual framework for the study to the data and analysis, and, in turn, to the conclusions. The conclusions are backed up by the evaluation findings."*

Evaluation Quality Tool: Source of the Items

Questions	Score	Source (inspired by/adapted from)
Conceptual Framing		
[1] Are the research/evaluation questions included in the report?	yes/no	ADS 201maa: Evaluation reports should adequately address all evaluation questions included in the SOW, or the evaluation questions subsequently revised and documented in consultation and agreement with USAID.
[2] Does the report include research/evaluation hypotheses?	yes/no/not applicable	BE^2, Checklist: Does the study outline a hypothesis?
[3] Are the research/evaluation questions appropriate for the intervention's conceptual framework (logframe/theory of change/results framework)?	yes/partial/no/ not applicable	BE^2, Checklist: Does the study pose an appropriate research question?
[4] Does the report acknowledge/draw upon existing country-specific research?	yes/partial/no	BE^2, Checklist: Does the study acknowledge existing research?
[5] Does the report explain the local context in sufficient detail?	yes/partial/no	USAID Evaluation Policy, page 8: Evaluation reports should include sufficient local and global contextual information so that the external validity and relevance of the evaluation can be assessed.
[6] Conceptual framing: Conclusion	adequate/not adequate	
[7] Conceptual framing: Notes/Justification		
Openness and Transparency		
[8] Is the report open about study limitations/weaknesses due to the methodology, sampling, data collection, etc.?	yes/partial/no	BE^2, page 17: The study should also clearly state the sample size.
[9] Is the report open about study limitations due to issues with the implementation of the intervention being evaluated?	yes/partial/no/ not applicable	BE^2, page 17: An important sign of quality is whether the author is being self-critical; being open about limitations.
[10] Does the report include alternative interpretations of the findings?	yes/no	BE^2, page 17: An important sign of quality is whether the author is being self-critical; being open about (...) alternative interpretations and pointing out inconsistencies with other results.
[11] Does the report present the complete analysis of data relevant for study questions?	yes/partial/no/ not applicable	BE^2, Checklist: Does the study present the raw data it analyses?
[12] Is the report open about potential biases due to the study team composition?	yes/partial/no	BE^2, Checklist: Does the researcher acknowledge their own subjectivity in the process of the research?
[13] Openness and transparency: Conclusion	adequate/not adequate	
[14] Openness and transparency: Notes/Justification		

Questions	Score	Source (inspired by/adapted from)
Robustness of Methodology		
[15] Is the methodology explained in sufficient detail?	yes/partial/no	ADS 201maa: Evaluation methodology should be explained in detail and sources of information properly identified.
[16] Is the methodology appropriate for answering posed study questions?	yes/partial/no	USAID Evaluation Policy, page 8: evaluation should principally consider the appropriateness of the evaluation design for answering the evaluation questions as well as balance cost, feasibility, and the level of rigor needed to inform specific decisions.
[17] Does the counterfactual meet standards of rigor?	yes/**no**/not applicable	USAID Evaluation Policy, page 3: Impact evaluations measure the change in a development outcome that is attributable to a defined intervention; impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change. (...) Performance evaluations encompass a broad range of evaluation methods. They often incorporate before-after comparisons, but generally lack a rigorously defined counterfactual.
[18] Is data triangulation described as part of methodology?	yes/partial/no/ not applicable	CASP, Qualitative Checklist: To what extent contradictory data are taken into account?
[19] Does the report mention steps to mitigate common threats to the integrity of the evaluation (such as non-equivalence at baseline, non-compliance, spillover, systematic attrition) or common biases (confounding bias, selection bias, experimenter bias, etc.)?	yes/partial/no/ not applicable	USAID Evaluation Policy, page 10: Evaluation reports that include the original statement of work, a full description of methodology (or methodologies) used, as well as the limitations in the inferences that can be drawn.
[20] Are sampling approach and sample size calculations presented in sufficient detail (to include, at a minimum, type of analysis, MDES, alpha and beta)?	yes/partial/no/ not applicable	JPAL's Running Randomized Evaluations, page 271: A power function relates power to its determinants: (1) level of significance, (2) MDE size, (3) the unexplained variance of the outcome of interest, (4) allocation fractions, (5) and the sample size.
[21] Is the sampling approach described in sufficient detail? (at a minimum, a rationale for the sample size and method of sample selection) and is it appropriate for the study objectives?	yes/partial/no/ not applicable	CASP, Qualitative Checklist: Recommended considerations about "If the researcher has explained how the participants were selected"; "If they explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study"; If there are any discussions around recruitment (e.g. why some people chose not to take part)".
[22] Robustness of methodology: Conclusion	adequate/not adequate	
[23] Robustness of methodology: Notes/Justification		

Questions	Score	Source (inspired by/adapted from)
Cultural Appropriateness		
[24] Does the report include documentation of local ethics review and/or US-based IRB approval/exemption status?	yes/no/not applicable	USAID Scientific Research Policy, page 6: Using research methods such as surveys, assessments, focus groups, polls and other quantitative and qualitative analytical techniques does not imply that the activity is research but, in many instances, the activity may still be called a "study" and subject to the policies outlined herein including review by an institutional review board (IRB) for human subjects' protections where required by regulation.
[25] Does the report list steps taken to ensure that study questions and methodology are informed by local stakeholders, are culturally relevant and appropriate?	yes/no	ADS 201sae: Is there reasonable assurance that the data collection methods being used do not produce systematically biased data.
[26] Does the report list steps to ensure that data collection tools were developed/adapted with participation of relevant local stakeholders and are culturally appropriate?	yes/partial/no	BE^2, page 20: For all research designs, it is important to consider the extent to which the measures/instruments/variables used in the study suit local contexts. The reviewer should note whether measures have been developed to suit the local context: does the study, for instance, merely translate into a local language or recognize that a test developed in a specific linguistic area may not be automatically suitable to a local context with translation or because of multiple socio-linguistic processes? The reviewer should also note whether local knowledge has been used effectively in the adaptation of measures to reflect resources relevant to the context; for example, are the instruments designed with support and recognition from the local community?
[27] Does the report list steps taken to validate findings/conclusions/recommendations with local stakeholders?	yes/no	EGRA Toolkit, 2nd edition, page 122: Results must be communicated to the appropriate audiences in a culturally and contextually suitable way in order to support understanding and action.
[28] Is the study informed by locally relevant stratifiers, such as political, social, ethnic, religious, geographical or sex/gender phenomena (including methodology, data collection and data analysis)?	yes/partial/no	BE^2, page 20: This includes the extent to which the analysis includes locally relevant social stratifiers (for example, socio-economic status, gender, rural-urban differences, etc.) and influences which may affect interpretation of results.
[29] Cultural appropriateness: Conclusion	adequate/not adequate	
[30] Cultural appropriateness: Notes/Justification		
Validity		
[31] Do indicators used in the evaluation serve as appropriate proxies for the construct or phenomenon being investigated?	yes/partial/no	BE^2, page 24: In the case of measurement validity, it is important to repeatedly consider whether or not the indicator chosen fully captures the concept being measured. Are there other dimensions of the central concept that are being ignored?
[32] Were the assessments conducted in such a way such that the results are generalizable to the population	yes/partial/no/ not applicable	StataCorp's Survey Data Reference Manual, page 3: In sample surveys, observations are selected through a random process, but different observations may have different probabilities of selection. Weights are equal to (or proportional to) the inverse of the probability of being sampled. Various post-sampling adjustments to

Questions	Score	Source (inspired by/adapted from)
of students reached through the activity?		the weights are sometimes made, as well. A weight of wj for the jth observation means, roughly speaking, that the jth observation represents wj elements in the population from which the sample was drawn. Omitting weights from the analysis results in estimates that may be biased, sometimes seriously so.
[33] Does the report allude to whether the study findings may have been biased by the activity of doing the study itself?	yes/no	BE^2, page 25: whether the findings could have been influenced by the process of research itself (ecological validity).
[34] Does the report address the external validity of findings?	yes/partial/no	BE^2, Checklist: To what extent is the study externally valid?
[35] Were all data collection tools piloted with representatives of target populations prior to beginning of the data collection?	yes/partial/no	EGRA Toolkit, 2nd edition, page 92: The students and schools selected for the pilot sample should be similar to the target population of the full study.
[36] Are confidence intervals reported around point estimates?	yes/no/not applicable	ADS 201sae: Has the margin of error been reported along with the data? (Only applicable to results obtained through statistical samples.)
[37] Are reported relationships tested for statistical significance and *p*-value reported?	yes/no/not applicable	What Works Clearinghouse Procedures and Standards, page 25: The WWC applies the Benjamini-Hochberg Correction for Multiple Comparisons only to statistically significant findings because nonsignificant findings will remain nonsignificant after correction. If the exact p-values are not available but effect sizes are available, the WWC converts the effect size to t-statistics and then obtains the corresponding p-values.
[38] Is treatment effect presented in terms of effect size?	yes/no/not applicable	What Works Clearinghouse Procedures and Standards, page 22: For all studies, the WWC records the study findings in the units reported by the study authors. In addition, the WWC computes and records the effect size associated with study findings on relevant outcome measures.
[39] Validity: Conclusion	adequate/not adequate	
[40] Validity: Notes/Justification		
Reliability		
[41] Does the report list steps taken to ensure that data were collected with a high degree of reliability?	yes/partial/no	ADS 201sae: Are data collection and analysis methods documented in writing and being used to ensure the same procedures are followed each time?
[42] Does the report provide statistics on inter-rater reliability of assessors during field data collection?	yes/no/not applicable	EGRA Toolkit 2nd edition, page 89: In addition to the assessor evaluation process during training, it is required that assessors continue to test the reliability and consistency among themselves (interrater reliability, or IRR) once they are in the field collecting data.
[43] Are data collection tools included in the report?	yes/partial/no	ADS 201mah: All data collection and analysis tools used, such as questionnaires, checklists, survey instruments, and discussion guides.
[44] Does the report adequately address missing data/non-response?	yes/partial/no	What Works Clearinghouse Procedures and Standards, page D.4: study must report the number of students (teachers, schools, etc.) who were designated as treatment and comparison group samples

Questions	Score	Source (inspired by/adapted from)
		and the proportion of the total sample (e.g., students, teachers, or schools in the treatment and comparison samples combined) with outcome data who were included in the impact analysis (i.e., response rates). Both overall attrition and attrition by treatment status must be reported.
[45] Reliability: Conclusion	adequate/not adequate	
[46] Reliability: Notes/Justification		
Cogency		
[47] Are all of the study questions, including sub-questions, answered?	yes/partial/no	ADS 201mah: Address all evaluation questions in the Statement of Work (SOW) or document approval by USAID for not addressing an evaluation question.
[48] Does the Executive Summary include answers to all of the study questions?	yes/no	ADS 201maa: The Executive Summary of an evaluation report should present a concise and accurate statement of the most critical elements of the report.
[49] Is the report written in a language accessible to the audiences for whom the report indicates it is written?	yes/no	USAID Evaluation Policy, page 10: USAID evaluations of all types will use sound social science methods and should include the following basic features: (...) Evaluation reports that are shared widely and in an accessible form with all partners and stakeholders, and with the general public.
[50] Are recommendations action-oriented, practical and specific?	yes/partial/no/ not applicable	ADS 201maa: If recommendations are included, they should be supported by a specific set of findings and should be action-oriented, practical, and specific.
[51] Is there a transparent connection between the study questions, findings from the data and the conclusions and recommendations, and is the report structured to present findings clearly and objectively?	yes/partial/no	E3 Sectoral Synthesis Checklist, question 32: Can a reader can follow a transparent path from findings to conclusions to recommendations?
[52] Are visuals in the report appropriate for helping non-technical audiences easily understand the study findings?	yes/partial/no	EGRA Toolkit 2nd edition, page 120: Data visualization must be used to facilitate understanding of the findings by general audiences. Visualizations are "standalone," such that the visual is interpretable without the audience needing to read extra text.
[53] Cogency: Conclusion	adequate/not adequate	
[54] Cogency: Notes/Justification		

Evaluation Quality Tool: Items by Evaluation Type

PRINCIPLE OF QUALITY	IMPACT EVALUATIONS	PERFORMANCE EVALUATIONS		STUDY	REVIEW RESULT	OVERALL CONCLUSION	NOTES/ JUSTIFICATION
		QUANTITATIVE	QUALITATIVE				
CONCEPTUAL FRAMING	Are the research/evaluation questions included in the report?				yes/no	adequate/not adequate	
	Does the report include research/evaluation hypotheses?				yes/no/not applicable		
	Are the research/evaluation questions appropriate for the intervention's conceptual framework (logframe/theory of change/results framework)?				yes/partial/no/not applicable		
	Does the report acknowledge/draw upon existing country-specific research?				yes/partial/no		
	Does the report explain the local context in sufficient detail?				yes/partial/no		
OPENNESS AND TRANSPARENCY	Is the report open about study limitations/weaknesses due to the methodology, sampling, data collection, etc?				yes/partial/no	adequate/not adequate	
	Is the report open about study limitations due to issues with the implementation of the intervention being evaluated?				yes/partial/no/not applicable		
	Does the report include alternative interpretations of the findings?				yes/no		
	Does the report present the complete analysis of data relevant for study questions?				yes/partial/no/not applicable		
	Is the report open about potential biases due to the study team composition?				yes/partial/no		
ROBUSTNESS OF METHODOLOGY	Is the methodology explained in sufficient detail?				yes/partial/no	adequate/not adequate	
	Is the methodology appropriate for answering posed study questions?				yes/partial/no		
	Does the counterfactual meet standards of rigor?				yes/no/not applicable		
		Is data triangulation described as part of methodology?			yes/partial/no/not applicable		
	Does the report mention steps to mitigate common threats to the integrity of the evaluation (such as non-equivalence at baseline, non-compliance, spillover, systematic attrition) or common biases (confounding bias, selection bias, experimenter bias, etc)?				yes/partial/no/not applicable		
	Are sampling approach and sample size calculations presented in sufficient detail (to include, at a minimum, type of analysis, MDES, alpha and beta)?				yes/partial/no/not applicable		

PRINCIPLE OF QUALITY	IMPACT EVALUATIONS	PERFORMANCE EVALUATIONS		STUDY	REVIEW RESULT	OVERALL CONCLUSION	NOTES/ JUSTIFICATION
		QUANTITATIVE	QUALITATIVE				
			Is the sampling approach described in sufficient detail? (at a minimum, a rationale for the sample size and method of sample selection) and is it appropriate for the study objectives?		yes/partial/no/not applicable		
CULTURAL APPROPRIATENESS	Does the report include documentation of local ethics review and/or US-based IRB approval/exemption status?				yes/no/not applicable	adequate/not adequate	
	Does the report list steps taken to ensure that study questions and methodology are informed by local stakeholders, are culturally relevant and appropriate?				yes/no		
	Does the report list steps to ensure that data collection tools were developed/adapted with participation of relevant local stakeholders and are culturally appropriate?				yes/partial/no		
	Does the report list steps taken to validate findings/conclusions/recommendations with local stakeholders?				yes/no		
	Is the study informed by locally relevant stratifiers, such as political, social, ethnic, religious, geographical or sex/gender phenomena (including methodology, data collection and data analysis)?				yes/partial/no		
VALIDITY	Do indicators used in the evaluation serve as appropriate proxies for the construct or phenomenon being investigated?				yes/partial/no	adequate/not adequate	
	Were the assessments conducted in such a way such that the results are generalizable to the population of students reached through the activity?				yes/partial/no/not applicable		
	Does the report allude to whether the study findings may have been biased by the activity of doing the study itself?				yes/no		
	Does the report address the external validity of findings?				yes/partial/no		
	Were all data collection tools piloted with representatives of target populations prior to beginning of the data collection?				yes/partial/no		
	Are confidence intervals reported around point estimates?				yes/no/not applicable		
	Are reported relationships tested for statistical significance and p-value reported?				yes/no/not applicable		
	Is treatment effect presented in terms of effect size?				yes/no/not applicable		
RELIABILITY	Does the report list steps taken to ensure that data were collected with a high degree of reliability?				yes/partial/no	adequate/not adequate	

PRINCIPLE OF QUALITY	IMPACT EVALUATIONS	PERFORMANCE EVALUATIONS		STUDY	REVIEW RESULT	OVERALL CONCLUSION	NOTES/ JUSTIFICATION
		QUANTITATIVE	QUALITATIVE				
	Does the report provide statistics on inter-rater reliability of assessors during field data collection?				yes/no/not applicable		
	Are data collection tools included in the report?				yes/partial/no		
	Does the report adequately address missing data/non-response?				yes/partial/no		
COGENCY	Are all of the study questions, including sub-questions, answered?				yes/no/not applicable	adequate/not adequate	
	Does the Executive Summary include answers to all of the study questions?				yes/no		
	Is the report written in a language accessible to the audiences for whom the report indicates it is written?				yes/no		
	Are recommendations action-oriented, practical and specific?				yes/partial/no/not applicable		
	Is there a transparent connection between the study questions, findings from the data and the conclusions and recommendations, and is the report structured to present findings clearly and objectively?				yes/partial/no		
	Are visuals in the report appropriate for helping non-technical audiences easily understand the study findings?				yes/partial/no		

Development of Evaluation Quality Tool: Reviewers' Feedback

Members of the education community generously shared their time and expertise to participate in a full-day reviewers' meeting at MSI's home office in Arlington VA, in which they discussed each item and question, answer options, and descriptors used in the evaluation quality tool. Thirty-six reviewers participated in the full-day reviewers' meeting, including 25 in person and 9 remotely.[60] For the discussion of the evaluation quality tool, they were put into four groups of eight or nine participants and provided with a set of questions for discussion.[61] During the review process, reviewers also shared feedback on the tool through the web platform developed for this study.

The key findings in this section represent the emerging themes in the reviewers' feedback. Following this piloting of the evaluation quality tool by the education community, the Office of Education has shown an interest in periodically repeating this exercise. This feedback will be used by the study team and the Office of Education to revise the evaluation quality tool for future evaluation quality assessments.

Scope of the Tool

Breadth. Reviewers' reactions were positive. They generally agreed on the usefulness of mapping the items to principles of quality, and some mentioned that adding principles of validity and reliability led to helpful reflections about the design of evaluations. Some reviewers mentioned that the value for money of the research design might be another desirable dimension to consider.

Length. Reviewers' reactions were mixed. Some thought that the 54 items made the tool too lengthy, that it took too long to administer the tool, and that the tool could be further simplified to narrow down the essential aspects of each principle of quality. others thought that the tool was too short, not capturing enough information about each principle of quality.

Applicability. Reviewers' reactions were mixed. Some thought it was too ambitious to have one tool applied to different types of evaluations (impact and performance) covering different research methods (quantitative and qualitative), while others found it helpful to have one tool focusing on the essential elements of the principles of quality that were common across evaluation types and research methods.

> "Two evaluations I read are different by purpose and design and analysis. To review them against the standardized criteria/principles is an interesting learning process. On one hand, we may be biased for or against the reports because of our own agreements or disagreements with the criteria. On the other, the quality of evaluation studies that were carried out are highly influenced by costs, time, experience of the team, and context. As I was reviewing these evaluation reports, I often asked myself what I would do in the situation, data tool development, data collection, data analysis, stakeholder involvement, interpretation and reporting, potential use of the evaluation, etc. I could not come up with specific or concrete answers in each of the situations because of many unknown context-related or resource-related or time-related challenges. Anyway, this is a very interesting learning experience for me to use the common, standardized and comprehensive list of criteria when reviewing USAID supported evaluation reports."
>
> - Expert reviewer

[60] This includes four MSI team members and two USAID representatives who volunteered to be reviewers for this study.
[61] A recent review found that with a relatively homogeneous population, and using a semi-structured guide, as few as three to six groups are likely to identify 90 percent of important themes. See Greg Guest, Emily Namey, and Kevin McKenna, "How many focus groups are enough? Building an evidence base for nonprobability sample sizes," *Field Methods* 29, no. 1 (2017): 3-22.

Items[62]

Match to principles of quality. Reviewers' reactions were positive. They perceived most items included in the study as relevant to the associated principles of quality. Some reviewers mentioned that the number of items under reliability should be expanded, especially to address steps to ensure consistency of results from repeated processing and analysis.

Match to evaluation types. Reviewers' reactions were positive. They perceived most items included in the study as relevant to the evaluation type to which they were applied. Some reviewers recommended expanding the number of items under validity associated with qualitative performance evaluations.

Ordering. Reviewers' reactions were negative. Several reviewers mentioned that ordering the items in the evaluation quality tool group by evaluation principle forced the reviewers to jump back and forth while going over the evaluation report. They recommended that items in the evaluation quality tool be reordered to better fit the flow of the report, and that items applicable only to certain evaluation types be made more explicit in the tool.

> "Perhaps, we may re-calibrate the criteria into 'common features' that should be seen for all evaluations and 'unique features' for specific types of evaluation, such as impact evaluation with quantitative methodology, performance evaluation to learn about fidelity of implementation by program design, etc."
>
> - Expert reviewer

Types. Some reviewers pointed out that the items might be capturing two separate constructs: one related to the presence of an element in the report (e.g., whether evaluation questions were included in the report) and the other an expert judgment about an element (e.g., whether the methodology used was appropriate for answering the evaluation questions). Reviewers suggested that the response options for the expert judgments be replaced with "high/mid/low" instead of "yes/partial/no."

Concerns

Scoring. Reviewers raised fears about evaluations being assigned an overall quality score. As mentioned during the orientation session and reiterated throughout the review process, this study did not produce overall scores for each evaluation, instead allowing for flexibility and a range of criteria depending on the questions being asked and other factors specific to the evaluation statement of work and the context-specific needs of the host government, for example.

Ramifications. Some reviewers expressed that USAID already provides a comprehensive set of evaluation guidance, checklists, and templates, and they were therefore concerned that the evaluation quality tool might become yet another requirement for USAID evaluation partners. Reviewers also noted that if items from the evaluation quality tool were to be incorporated in a future procurement, USAID would need to assess the time and cost implications of improving evaluation quality to meet these standards. As one reviewer mentioned, to ensure a realistic timeframe, when commissioning the evaluation USAID would need to consider an illustrative timeline showing that a local ethics review could take nine months, for example.

As another reviewer mentioned concerning the impact on cost, "If the intervention is smallish or very specific (or is a smallish and very specific section of a bigger project) then there are also economy and

[62] Reviewers provided feedback on each individual item, and while this will be used by the Office of Education and the study team to review the tool, it was too granular to be included in the emerging themes about the overall tool.

efficiency concerns. The kinds of detail and thoroughness listed here are admirable, but in principle could be too costly for something that is a very small and/or very specific little intervention, where a 'lighter' (though not less rigorous) approach might be justified."

Overall Exercise

Repetition. During the reviewers' meeting, many reviewers supported the idea that the process used for this evaluation quality assessment be repeated on a periodic basis, for example annually. Reviewers mentioned that this process provided an opportunity for experts to read each other's evaluations, which led to constructive discussions about quality standards and the subject matter of the reviewed studies.

Main Takeaways

1. Reviewers voiced interest in participating in a future round of evaluation quality reviews.
2. In general, reviewers expressed satisfaction with the breadth of the tool, based on the principles of quality recommended by BE^2, and agreed with the mapping of individual items to the seven principles of quality. Reviewers offered mixed feedback on the appropriate length for the tool, with some thinking that 54 items were too few and others too many; this may suggest that between 50 and 60 items is an adequate length. Some items captured the presence of an element, while others captured expert judgments about an element. A possibility is to revise both the evaluation quality tool and its complementary background evaluation tool, moving items that capture the presence of an element to the supporting tool, and allowing the evaluation quality tool to focus only on expert judgments.
3. Reviewers offered mixed feedback on the benefits of having one tool to assess the quality of different evaluation types (impact and performance) and research methods (quantitative and qualitative), with some arguing that dedicated tools would be more precise, while others believed that a general tool was more efficient. Given the existence of other dedicated evidence rating systems, this tool's contribution may lie in its versatility. For the items that were not common across evaluations types, reviewers generally agreed on their mapping to evaluation types.
4. Reviewers were dissatisfied with the ordering of items in the tool, which were grouped around principles of quality, so future iterations of the tool may reorder the items to follow the outline in the USAID evaluation report template.
5. Reviewers generally agreed that the tool should not be used to produce a composite score about overall evaluation quality, and that the scoring of the adequacy of the principles of quality should be relative to the evaluation type. They also mentioned other circumstances to consider, such as adding value for money.

ANNEX 5: INFORMATION ABOUT EVALUATIONS ASSESSMENT RESULTS

Context

Geographic coverage of the evaluation	Percent
district (n=8)	8.7%
sub-national (n=30)	32.6%
national (n=47)	51.1%
multi-country (n=7)	7.6%

Scale of the evaluated program	Percent
pilot (n=25)	27.2%
full intervention (n=61)	66.3%
not applicable (n=6)	6.5%

Evaluation type	Percent
impact, experimental (n=13)	14.1%
impact, quasi-experimental (n=12)	13.0%
impact, non-experimental* (n=2)	2.2%
performance, qualitative (n=46)	50%
performance, quantitative (n=13)	14.1%
study (n=6)	6.5%

* See question about rigor of counterfactual in the evaluation quality assessment

Stage of evaluation	Percent
baseline (n=1)	1.1%
mid-term or midline (n=33)	35.9%
final or endline (n=52)	56.5%
not applicable (n=6)	6.5%

Report includes information about intervention dosage at the beneficiary level	Percent
no (n=20)	21.7%
not applicable (n=6)	6.5%
yes, for some beneficiaries (n=19)	20.7%
yes, for all beneficiaries (n=47)	51.1%

Evaluation respondents (multiple responses allowed)	Percent
primary grade students (n=28)	30.4%
secondary grade students (n=12)	13.0%
vocational school or tertiary level students (n=20)	21.7%
non-formal education or alternative education learners (n=2)	2.2%
parents (n=18)	19.6%
early childhood educators or master trainers (n=29)	31.5%
upper primary or secondary school educators or master trainers (n=26)	28.3%
tertiary or vocational instructors or master trainers (n=24)	26.1%
government officials or administrators (n=51)	55.4%
entrepreneurs (n=10)	10.9%

Learning assessments included in the evaluation (multiple responses allowed)	Percent
none (n=62)	67.4%
early childhood assessment (n=2)	2.2%
early grade reading (n=17)	18.5%
early grade math (n=10)	10.9%
vocational skills (n=5)	5.4%
soft skills or social-emotional skills (n=4)	4.3%

Other assessments included in the evaluation (multiple responses allowed)	Percent
none (n=34)	37.0%
assessment of school management or leadership (n=16)	17.4%
assessment of teacher knowledge or practice (n=19)	20.7%
assessment of teacher wellbeing or motivation (n=7)	7.6%
assessment of learning environment, including safe learning environment or GBV (n=20)	21.7%
assessment of learners' wellbeing (n=17)	18.5%
assessment of community or parents or caregivers (n=21)	22.8%

Conceptual Framing

Report includes project's logic model/results framework/theory of change	
no (n=29)	31.5%
yes (n=57)	62.0%
not applicable (n=6)	6.5%

Report includes a description of the intervention?	
no (n=1)	1.1%
yes, but little detail (n=7)	7.6%
yes, in detail (n=78)	84.8%
not applicable (n=6)	6.5%

Evaluation questions includes explicit exploration of cross-cutting topics (multiple responses allowed)	
no (n=41)	44.6%
yes, gender (n=32)	34.8%
yes, disability (n=7)	7.6%
yes, ICT in education (n=4)	4.3%
yes, innovative finance (n=2)	2.2%
yes, scale and sustainability (n=38)	41.3%

Openness and Transparency

Fidelity of implementation from M&E system included in the data analysis	
no (n=78)	84.8%
yes, data from the assessment (n=8)	8.7%
not applicable (n=6)	6.5%

Report includes information about project targets?	
no (n=24)	26.1%
yes (n=62)	67.4%
not applicable (n=6)	6.5%

Robustness of the Methodology

Project's monitoring data included in the evaluation data analysis?	
no (n=46)	50.0%
yes (n=40)	43.5%
not applicable (n=6)	6.5%

Cultural Appropriateness

Staffing of evaluation team (multiple responses allowed)	
international evaluation specialists (n=56)	60.9%
local evaluation specialists (n=58)	63.0%
representative of the implementation team (n=15)	16.3%
USAID staff (n=3)	3.3%
other international stakeholders (n=4)	4.3%
other local stakeholders (n=18)	19.6%
not reported (n=17)	18.5%

Results disaggregated by subgroups (multiple responses allowed)	
none (n=19)	20.7%
gender (n=67)	72.8%
socio-economic status (n=6)	6.5%
ethnic or linguistic group (n=12)	13.0%
disability status (n=3)	3.3%
grade level (n=26)	28.3%

Cogency

Information about audience for the report	
no (n=22)	23.9%
yes (n=70)	76.1%

Report explains how findings will be used	
no (n=5)	5.4%
yes, but little detail (n=29)	31.5%
yes, in detail (n=58)	63.0%

ANNEX 6: EVALUATION QUALITY ASSESSMENT RESULTS

Results by Evaluation Type

Conceptual Framing

Study questions included

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=18)	33.3%	15.4%	8.7%	50.0%	19.6%
yes (n=74)	66.7%	84.6%	91.3%	50.0%	80.4%

Cramer's V = 0.3359

Study hypotheses included

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (NA)	total (n = 86)
no (n=60)	44.4%	61.5%	87.0%	---	69.8%
yes (n=26)	55.6%	38.5%	13.0%	---	30.2%

Cramer's V = 0.4186
NA = study (6)

Study questions appropriate given the intervention's conceptual framework

	impact (n = 18)	perf. quant. (n = 11)	perf. qual. (n = 42)	study (NA)	total (n = 71)
no (n=2)	5.6%	0.0%	2.4%	---	2.8%
partial (n=21)	5.6%	18.2%	42.9%	---	29.6%
yes (n=48)	88.9%	81.8%	54.8%	---	67.6%

Cramer's V = 0.2637
NA = impact (9), perf. quant. (2), perf. qual. (4), study (6)

Study acknowledges/draws upon existing country-specific research

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=25)	25.9%	30.8%	30.4%	0.0%	27.2%
partial (n=35)	44.4%	30.8%	39.1%	16.7%	38.0%
yes (n=32)	29.6%	38.5%	30.4%	83.3%	34.8%

Cramer's V = 0.2041

Local context provided allows non-experts to appreciate relevance of the study

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=18)	22.2%	23.1%	19.6%	0.0%	19.6%
partial (n=32)	25.9%	46.2%	39.1%	16.7%	34.8%
yes (n=42)	51.9%	30.8%	41.3%	83.3%	45.7%

Cramer's V = 0.1832

Conceptual Framing: Conclusion

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
adequate (n=62)	70.4%	76.9%	60.9%	83.3%	67.4%
not adequate (n=30)	29.6%	23.1%	39.1%	16.7%	32.6%

Cramer's V = 0.1557

Openness and Transparency

Open about limitations to implementing the study

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=11)	29.6%	0.0%	4.3%	16.7%	12.0%
partial (n=43)	29.6%	69.2%	52.2%	33.3%	46.7%
yes (n=38)	40.7%	30.8%	43.5%	50.0%	41.3%

Cramer's V = 0.2863

Open about limitations to implementing the intervention

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 45)	study (n = 6)	total (n = 85)
no (n=17)	14.8%	53.8%	13.3%	---	20.0%
partial (n=31)	25.9%	38.5%	42.2%	---	36.5%
yes (n=37)	59.3%	7.7%	44.4%	---	43.5%

Cramer's V = 0.3013
NA = perf. qual. (1), study (6)

Alternative interpretations of the findings included

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=72)	59.3%	92.3%	87.0%	66.7%	78.3%
yes (n=20)	40.7%	7.7%	13.0%	33.3%	21.7%

Cramer's V = 0.3256

Comprehensive analysis of the data relevant for study questions included

	impact (n = 23)	perf. quant. (n = 13)	perf. qual. (n = 45)	study (n = 5)	total (n = 86)
no (n=10)	8.7%	0.0%	17.8%	0.0%	11.6%
partial (n=36)	13.0%	69.2%	42.2%	100.0%	41.9%
yes (n=40)	78.3%	30.8%	40.0%	0.0%	46.5%

Cramer's V = 0.3662
NA = impact (4), perf. qual. (1), study (1)

Open about potential biases due to the study team composition

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=62)	66.7%	76.9%	63.0%	83.3%	67.4%
partial (n=20)	25.9%	7.7%	23.9%	16.7%	21.7%
yes (n=10)	7.4%	15.4%	13.0%	0.0%	10.9%

Cramer's V = 0.1378

Openness and Transparency: Conclusion

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
adequate (n=45)	59.3%	46.2%	45.7%	33.3%	48.9%
not adequate (n=47)	40.7%	53.8%	54.3%	66.7%	51.1%

Cramer's V = 0.1465

Robustness of the Methodology

Methodology explained in detail

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=5)	0.0%	15.4%	6.5%	0.0%	5.4%
partial (n=41)	29.6%	61.5%	52.2%	16.7%	44.6%
yes (n=46)	70.4%	23.1%	41.3%	83.3%	50.0%

Cramer's V = 0.2783

Methodology appropriate for the study

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=6)	3.7%	15.4%	6.5%	0.0%	6.5%
partial (n=40)	22.2%	53.8%	50.0%	66.7%	43.5%
yes (n=46)	74.1%	30.8%	43.5%	33.3%	50.0%

Cramer's V = 0.2491

Counterfactual meets standards of rigor

	impact (n = 27)	perf. quant. (NA)	perf. qual. (NA)	study (NA)	total (n = 27)
no (n=6)	22.2%	---	---	---	22.2%
yes (n=21)	77.8%	---	---	---	77.8%

Cramer's V = NA
NA: perf. quant. (13), perf. qual. (46), study (6)

Data triangulation described as part of methodology

	impact (NA)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 65)
no (n=9)	---	15.4%	13.0%	16.7%	13.8%
partial (n=23)	---	30.8%	41.3%	0.0%	35.4%
yes (n=33)	---	53.8%	45.7%	83.3%	50.8%

Cramer's V = 0.1808
NA: impact (27)

Addressed internal validity, either threats to inference or common biases

	impact (n = 26)	perf. quant. (n = 13)	perf. qual. (n = 41)	study (n = 2)	total (n = 82)
no (n=28)	7.7%	46.2%	46.3%	50.0%	34.1%
partial (n=38)	57.7%	53.8%	39.0%	0.0%	46.3%
yes (n=16)	34.6%	0.0%	14.6%	50.0%	19.5%

Cramer's V = 0.3231
NA: impact (1), perf. qual. (5), study (4)

Described sampling approach and parameters used to compute sample size

	impact (n = 26)	perf. quant. (n = 13)	perf. qual. (NA)	study (NA)	total (n = 40)
no (n=9)	14.8%	38.5%	---	---	22.5%
partial (n=11)	29.6%	23.1%	---	---	27.5%
yes (n=20)	55.6%	38.5%	---	---	50.0%

Cramer's V = 0.2660
NA: perf. qual. (46), study (6)

Described sampling approach to collect qualitative data

	impact (NA)	perf. quant. (NA)	perf. qual. (n = 44)	study (n = 5)	total (n = 49)
no (n=11)	---	---	25.0%	0.0%	22.4%
partial (n=23)	---	---	47.7%	40.0%	46.9%
yes (n=15)	---	---	27.3%	60.0%	30.6%

Cramer's V = 0.2424
NA: impact (27), perf. quant. (13), perf. qual. (2), study (1)

Robustness of Methodology: Conclusion

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
adequate (n=60)	81.5%	69.2%	54.3%	66.7%	65.2%
not adequate (n=32)	18.5%	30.8%	45.7%	33.3%	34.8%

Cramer's V = 0.2476

Cultural Appropriateness

Included documentation from ethics review for approval/exemption status

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 44)	study (n = 5)	total (n = 89)
no (n=79)	81.5%	100%	93.2%	60.0%	88.8%
yes (n=10)	18.5%	0.0%	6.8%	40.0%	11.2%

Cramer's V = 0.3015
NA: perf. qual. (2), study (1)

Study questions and methodology informed by local stakeholders

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=73)	77.8%	84.6%	84.8%	33.3%	79.3%
yes (n=19)	22.2%	15.4%	15.2%	66.7%	20.7%

Cramer's V = 0.3100

Data collection tools developed with participation of local stakeholders

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=56)	66.7%	69.2%	56.5%	50.0%	60.9%
partial (n=22)	18.5%	23.1%	26.1%	33.3%	23.9%
yes (n=14)	14.8%	7.7%	17.4%	16.7%	15.2%

Cramer's V = 0.1000

Findings/conclusions/recommendations validated with local stakeholders

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=81)	100%	84.6%	80.4%	100.0%	88.0%
yes (n=11)	0.0%	15.4%	19.6%	0.0%	12.0%

Cramer's V = 0.2789

Findings disaggregated by locally relevant stratifiers

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=15)	11.1%	23.1%	19.6%	0.0%	16.3%
partial (n=46)	33.3%	53.8%	56.5%	66.7%	50.0%
yes (n=31)	55.6%	23.1%	23.9%	33.3%	33.7%

Cramer's V = 0.2322

Cultural Appropriateness: Conclusion

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
adequate (n=27)	25.9%	23.1%	26.1%	83.3%	29.3%
not adequate (n=65)	74.1%	76.9%	73.9%	16.7%	70.7%

Cramer's V = 0.3140

Validity

Addressed construct validity of the assessment tools

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (NA)	study (NA)	total (n = 40)
no (n=0)	0%	0%	---	---	0%
partial (n=6)	7.4%	30.8%	---	---	15.0%
yes (n=34)	92.6%	69.2%	---	---	85.0%

Cramer's V = -0.3064
NA: perf. qual. (46), study (6)

Addressed the external validity of findings from the sample to population

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (NA)	study (NA)	total (n = 40)
no (n=12)	11.1%	69.2%	---	---	30.0%
partial (n=8)	22.2%	15.4%	---	---	20.0%
yes (n=20)	66.7%	15.4%	---	---	50.0%

Cramer's V = 0.3064
NA: perf. qual. (46), study (6)

Addressed ecological validity of findings

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=80)	96.3%	84.6%	82.6%	83.3%	87.0%
yes (n=12)	3.7%	15.4%	17.4%	16.7%	13.0%

Cramer's V = 0.1798

Addressed the external validity of findings to other contexts

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=53)	37.0%	76.9%	67.4%	33.3%	57.6%
partial (n=24)	33.3%	23.1%	26.1%	0.0%	26.1%
yes (n=15)	29.6%	0.0%	6.5%	66.7%	16.3%

Cramer's V = 0.3570

Data collection tools piloted with representatives of target populations

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=56)	44.4%	38.5%	76.1%	66.7%	60.9%
partial (n=23)	33.3%	38.5%	17.4%	16.7%	25.0%
yes (n=13)	22.2%	23.1%	6.5%	16.7%	14.1%

Cramer's V = 0.2431

Confidence intervals reported around point estimates

	impact (n = 27)	perf. quant. (n = 7)	perf. qual. (NA)	study (NA)	total (n = 34)
no (n=18)	63.0%	14.3%	---	---	52.9%
yes (n=16)	37.0%	85.7%	---	---	47.1%

Cramer's V = 0.3943
NA: perf. quant. (6), perf. qual. (46), study (6)

Relationships tested for statistical significance and *p*-value reported

	impact (n = 27)	perf. quant. (n = 7)	perf. qual. (NA)	study (NA)	total (n = 34)
yes (n=34)	100%	100%	---	---	100%

Cramer's V = NA
NA: perf. quant. (6), perf. qual. (46), study (6)

Treatment effects presented in terms of effect sizes

	impact (n = 26)	perf. quant. (n = 7)	perf. qual. (NA)	study (NA)	total (n = 33)
no (n=14)	38.5%	57.1%	---	---	42.4%
yes (n=19)	61.5%	42.9%	---	---	57.6%

Cramer's V = -0.1545
NA: impact (1), perf. quant. (6), perf. qual. (46), study (6)

Validity: Conclusion

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
adequate (n=38)	66.7%	38.5%	26.1%	50.0%	41.3%
not adequate (n=54)	33.3%	61.5%	73.9%	50.0%	58.7%

Cramer's V = 0.3580

Reliability

Steps taken to ensure that data were reliably collected

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=41)	40.7%	30.8%	54.3%	16.7%	44.6%
partial (n=35)	40.7%	53.8%	28.3%	66.7%	38.0%
yes (n=16)	18.5%	15.4%	17.4%	16.7%	17.4%

Cramer's V = 0.1819

Inter-rater reliability statistics of assessors' fieldwork provided

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (NA)	study (NA)	total (n = 40)
no (n=36)	92.6%	84.6%	---	---	90.0%
yes (n=4)	7.4%	15.4%	---	---	10.0%

Cramer's V = 0.1245
NA: perf. qual. (46), study (6)

Data collection tools included in annex

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=22)	33.3%	30.8%	13.0%	50.0%	23.9%
partial (n=2)	7.4%	0.0%	0.0%	0.0%	2.2%
yes (n=68)	59.3%	69.2%	87.0%	50.0%	73.9%

Cramer's V = 0.2596

Target and actual sample sizes reported and non-responses bias discussed

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=42)	37.0%	38.5%	54.3%	33.3%	45.7%
partial (n=40)	37.0%	61.5%	43.5%	33.3%	43.5%
yes (n=10)	25.9%	0.0%	2.2%	33.3%	10.9%

Cramer's V = 0.2944

Reliability: Conclusion

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
adequate (n=34)	37.0%	23.1%	43.5%	16.7%	37.0%
not adequate (n=58)	63.0%	76.9%	56.5%	83.3%	63.0%

Cramer's V = 0.1798

Cogency

Answers to all study questions, including sub-questions, included

	impact (n = 18)	perf. quant. (n = 11)	perf. qual. (n = 42)	study (n = 3)	total (n = 74)
no (n=14)	11.1%	18.2%	21.4%	33.3%	18.9%
yes (n=60)	88.9%	81.8%	78.6%	66.7%	81.1%

Cramer's V = 0.1324

NA: impact (9), perf. quant. (2), perf. qual. (4), study (3)

Answers to all study questions included in the Executive Summary

	impact (n = 18)	perf. quant. (n = 11)	perf. qual. (n = 42)	study (n = 3)	total (n = 74)
no (n=21)	11.1%	45.5%	28.6%	66.7%	28.4%
yes (n=53)	88.9%	54.5%	71.4%	33.3%	71.6%

Cramer's V = 0.2937

NA: impact (9), perf. quant. (2), perf. qual. (4), study (3)

Written in a language adequate to its stated audience

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=7)	11.1%	23.1%	2.2%	0.0%	7.6%
yes (n=85)	88.9%	76.9%	97.8%	100.0%	92.4%

Cramer's V = 0.2821

Recommendations are action-oriented, practical and specific

	impact (n = 21)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 86)
no (n=6)	14.3%	7.7%	4.3%	0.0%	7.0%
partial (n=33)	38.1%	46.2%	37.0%	33.3%	38.4%
yes (n=47)	47.6%	46.2%	58.7%	66.7%	54.7%

Cramer's V = 0.1410

NA: impact (6)

Connection between study questions, findings, conclusions and recommendations

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=5)	3.7%	0.0%	8.7%	0.0%	5.4%
partial (n=43)	48.1%	38.5%	47.8%	50.0%	46.7%
yes (n=44)	48.1%	61.5%	43.5%	50.0%	47.8%

Cramer's V = 0.1279

Visuals are helpful for a non-technical audience to understand the findings

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
no (n=11)	7.4%	7.7%	17.4%	0.0%	12.0%
partial (n=40)	37.0%	46.2%	45.7%	50.0%	43.5%
yes (n=41)	55.6%	46.2%	37.0%	50.0%	44.6%

Cramer's V = 0.1526

Cogency: Conclusion

	impact (n = 27)	perf. quant. (n = 13)	perf. qual. (n = 46)	study (n = 6)	total (n = 92)
adequate (n=69)	74.1%	76.9%	73.9%	83.3%	75.0%
not adequate (n=23)	25.9%	23.1%	26.1%	16.7%	25.0%

Cramer's V = 0.0561

Results by Country Income Level

Conceptual Framing

Study questions included

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=18)	26.7%	19.6%	6.3%	19.6%
yes (n=74)	73.3%	80.4%	93.8%	80.4%

Cramer's V = 0.1733

Study hypotheses included

	Low (n = 26)	Lower-Middle (n = 44)	Upper-Middle (n = 16)	Total (n = 86)
no (n=60)	65.4%	65.9%	87.5%	69.8%
yes (n=26)	34.6%	34.1%	12.5%	30.2%

Cramer's V = 0.1847
NA = Low (4), Lower-Middle (2)

Study questions appropriate given the intervention's conceptual framework

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=2)	0.0%	0.0%	13.3%	2.8%
partial (n=21)	20.0%	36.1%	26.7%	29.6%
yes (n=48)	80.0%	63.9%	60.0%	67.6%

Cramer's V = 0.2560

Study acknowledges/draws upon existing country-specific research

	Low (n = 20)	Lower-Middle (n = 36)	Upper-Middle (n = 15)	Total (n = 71)
no (n=25)	26.7%	21.7%	43.8%	27.2%
partial (n=35)	30.0%	43.5%	37.5%	38.0%
yes (n=32)	43.3%	34.8%	18.8%	34.8%

Cramer's V = 0.1616
NA = Low (10), (Lower- Middle (10), Upper-Middle (1)

Local context provided allows non-experts to appreciate relevance of the study

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=18)	10.0%	17.4%	43.8%	19.6%
partial (n=32)	40.0%	34.8%	25.0%	34.8%
yes (n=42)	50.0%	47.8%	31.3%	45.7%

Cramer's V = 0.2068

Conceptual Framing: Conclusion

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
adequate (n=62)	60.0%	78.3%	50.0%	67.4%
not adequate (n=30)	40.0%	21.7%	50.0%	32.6%

Cramer's V = 0.2427

Openness and Transparency

Open about limitations to implementing the study

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=11)	16.7%	13.0%	0.0%	12.0%
partial (n=43)	53.3%	45.7%	37.5%	46.7%
yes (n=38)	30.0%	41.3%	62.5%	41.3%

Cramer's V = 0.1770

Open about limitations to implementing the intervention

	Low (n = 26)	Lower-Middle (n = 44)	Upper-Middle (n = 15)	Total (n = 85)
no (n=17)	15.4%	22.7%	20.0%	20.0%
partial (n=31)	34.6%	36.4%	40.0%	36.5%
yes (n=37)	50.0%	40.9%	40.0%	43.5%

Cramer's V = 0.0719
NA = Low (4), Lower-Middle (2), Upper-Middle (1)

Alternative interpretations of the findings included

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=72)	76.7%	73.9%	93.8%	78.3%
yes (n=20)	23.3%	26.1%	6.3%	21.7%

Cramer's V = 0.1748

Comprehensive analysis of the data relevant for study questions included

	Low (n = 27)	Lower-Middle (n = 44)	Upper-Middle (n = 15)	Total (n = 86)
no (n=10)	14.8%	6.8%	20.0%	11.6%
partial (n=36)	40.7%	50.0%	20.0%	41.9%
yes (n=40)	44.4%	43.2%	60.0%	46.5%

Cramer's V = 0.1730
NA = Low (3), Lower-Middle (2), Upper-Middle (1)

Open about potential biases due to the study team composition

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=62)	76.7%	69.6%	43.8%	67.4%
partial (n=20)	16.7%	23.9%	25.0%	21.7%
yes (n=10)	6.7%	6.5%	31.3%	10.9%

Cramer's V = 0.2293

Openness and Transparency: Conclusion

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
adequate (n=45)	40.0%	54.3%	50.0%	48.9%
not adequate (n=47)	60.0%	45.7%	50.0%	51.1%

Cramer's V = 0.1279

Robustness of the Methodology

Methodology explained in detail

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=5)	3.3%	4.3%	12.5%	5.4%
partial (n=41)	46.7%	45.7%	37.5%	44.6%
yes (n=46)	50.0%	50.0%	50.0%	50.0%

Cramer's V = 0.1052

Methodology appropriate for the study

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=6)	10.0%	6.5%	0.0%	6.5%
partial (n=40)	43.3%	43.5%	43.8%	43.5%
yes (n=46)	46.7%	50.0%	56.3%	50.0%

Cramer's V = 0.0987

Counterfactual meets standards of rigor

	Low (n = 11)	Lower-Middle (n = 15)	Upper-Middle (n = 1)	Total (n = 27)
no (n=6)	36.4%	6.7%	100%	22.2%
yes (n=21)	63.6%	93.3%	0.0%	77.8%

Cramer's V = 0.5045
NA = Low (19), Lower-Middle (31), Upper-Middle (15)

Data triangulation described as part of methodology

	Low (n = 19)	Lower-Middle (n = 31)	Upper-Middle (n = 15)	Total (n = 65)
no (n=9)	21.1%	12.9%	6.7%	13.8%
partial (n=23)	31.6%	38.7%	33.3%	35.4%
yes (n=33)	47.4%	48.4%	60.0%	50.8%

Cramer's V = 0.1181
NA = Low (11), Lower-Middle (15), Upper-Middle (1)

Addressed internal validity, either threats to inference or common biases

	Low (n = 24)	Lower-Middle (n = 42)	Upper-Middle (n = 16)	Total (n = 82)
no (n=28)	37.5%	26.2%	50.0%	34.1%
partial (n=38)	45.8%	52.4%	31.3%	46.3%
yes (n=16)	16.7%	21.4%	18.8%	19.5%

Cramer's V = 0.1426
NA = Low (6), Lower-Middle (4),

Described sampling approach and parameters used to compute sample size

	Low (n = 14)	Lower-Middle (n = 24)	Upper-Middle (n = 2)	Total (n = 36)
no (n=9)	21.4%	25.0%	0.0%	22.5%
partial (n=11)	21.4%	29.2%	50.0%	27.5%
yes (n=20)	57.1%	45.8%	50.0%	50.0%

Cramer's V = 0.1289
NA = Low (16), Lower-Middle (22), Upper-Middle (14)

Described sampling approach to collect qualitative data

	Low (n = 14)	Lower-Middle (n = 21)	Upper-Middle (n = 14)	Total (n = 49)
no (n=11)	14.3%	28.6%	21.4%	22.4%
partial (n=23)	50.0%	42.9%	50.0%	46.9%
yes (n=15)	35.7%	28.6%	28.6%	30.6%

Cramer's V = 0.1045
NA = Low (16), Lower-Middle (25), Upper-Middle (2)

Robustness of Methodology: Conclusion

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
adequate (n=60)	63.3%	67.4%	62.5%	65.2%
not adequate (n=32)	36.7%	32.6%	37.5%	34.8%

Cramer's V = 0.0460

Cultural Appropriateness

Included documentation from ethics review for approval/exemption status

	Low (n = 29)	Lower-Middle (n = 45)	Upper-Middle (n = 15)	Total (n = 89)
no (n=79)	82.8%	88.9%	100%	88.8%
yes (n=10)	17.2%	11.1%	0.0%	11.2%

Cramer's V = 0.1820
NA = Low (1), Lower-Middle (1), Upper-Middle (1)

Study questions and methodology informed by local stakeholders

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=73)	80.0%	80.4%	75.0%	79.3%
yes (n=19)	20.0%	19.6%	25.0%	20.7%

Cramer's V = 0.0495

Data collection tools developed with participation of local stakeholders

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=56)	76.7%	52.2%	56.3%	60.9%
partial (n=22)	16.7%	30.4%	18.8%	23.9%
yes (n=14)	6.7%	17.4%	25.0%	15.2%

Cramer's V = 0.1821

Findings/conclusions/recommendations validated with local stakeholders

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=81)	93.3%	89.1%	75.0%	88.0%
yes (n=11)	6.7%	10.9%	25.0%	12.0%

Cramer's V = 0.1932

Findings disaggregated by locally relevant stratifiers

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=15)	23.3%	15.2%	6.3%	16.3%
partial (n=46)	56.7%	47.8%	43.8%	50.0%
yes (n=31)	20.0%	37.0%	50.0%	33.7%

Cramer's V = 0.1721

Cultural Appropriateness: Conclusion

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
adequate (n=27)	23.3%	30.4%	37.5%	29.3%
not adequate (n=65)	76.7%	69.6%	62.5%	70.7%

Cramer's V = 0.1075

Validity

Addressed construct validity of the assessment tools

	Low (n = 14)	Lower-Middle (n = 24)	Upper-Middle (n = 2)	Total (n = 40)
no (n=0)	%	%	%	%
partial (n=6)	7.1%	20.8%	0.0%	15.0%
yes (n=34)	92.9%	79.2%	100%	85.0%

Cramer's V = 0.2044

NA = Low (16), Lower-Middle (22), Upper-Middle (14)

Addressed the external validity of findings from the sample to population

	Low (n = 14)	Lower-Middle (n = 24)	Upper-Middle (n = 2)	Total (n = 40)
no (n=12)	14.3%	37.5%	50.0%	30.0%
partial (n=8)	35.7%	8.3%	50.0%	20.0%
yes (n=20)	50.0%	54.2%	0.0%	50.0%

Cramer's V = 0.3003
NA = Low (16), Lower-Middle (22), Upper-Middle (14)

Addressed ecological validity of findings

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=80)	86.7%	91.3%	75.0%	87.0%
yes (n=12)	13.3%	8.7%	25.0%	13.0%

Cramer's V = 0.1740

Addressed the external validity of findings to other contexts

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=53)	63.3%	52.2%	62.5%	57.6%
partial (n=24)	16.7%	30.4%	31.3%	26.1%
yes (n=15)	20.0%	17.4%	6.3%	16.3%

Cramer's V = 0.1332

Data collection tools piloted with representatives of target populations

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=56)	73.3%	52.2%	62.5%	60.9%
partial (n=23)	13.3%	34.8%	18.8%	25.0%
yes (n=13)	13.3%	13.0%	18.8%	14.1%

Cramer's V = 0.1694

Confidence intervals reported around point estimates

	Low (n = 14)	Lower-Middle (n = 19)	Upper-Middle (n = 1)	Total (n = 34)
no (n=18)	57.1%	52.6%	0.0%	52.9%
yes (n=16)	42.9%	47.4%	100%	47.1%

Cramer's V = 0.1898
NA = Low (16), Lower-Middle (27), Upper-Middle (15)

Relationships tested for statistical significance and *p*-value reported

	Low (n = 14)	Lower-Middle (n = 19)	Upper-Middle (n = 1)	Total (n = 34)
yes (n=34)	100%	100%	100%	100%

Cramer's V = NA
NA = Low (16), Lower-Middle (27), Upper-Middle (15)

Treatment effects presented in terms of effect sizes

	Low (n = 14)	Lower-Middle (n = 18)	Upper-Middle (n = 1)	Total (n = 33)
no (n=14)	28.6%	55.6%	0.0%	42.4%
yes (n=19)	71.4%	44.4%	100%	57.6%

Cramer's V = 0.3069
NA = Low (16), Lower-Middle (28), Upper-Middle (15)

Validity: Conclusion

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
adequate (n=38)	33.3%	45.7%	43.8%	41.3%
not adequate (n=54)	66.7%	54.3%	56.3%	58.7%

Cramer's V = 0.1135

Reliability

Steps taken to ensure that data were reliably collected

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=41)	46.7%	43.5%	43.8%	44.6%
partial (n=35)	33.3%	41.3%	37.5%	38.0%
yes (n=16)	20.0%	15.2%	18.8%	17.4%

Cramer's V = 0.0575

Inter-rater reliability statistics of assessors' fieldwork provided

	Low (n = 14)	Lower-Middle (n = 24)	Upper-Middle (n = 2)	Total (n = 40)
no (n=36)	92.9%	87.5%	100%	90.0%
yes (n=4)	7.1%	12.5%	0.0%	10.0%

Cramer's V = 0.1136

NA = Low (16), Lower-Middle (22), Upper-Middle (14)

Data collection tools included in annex

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=22)	30.0%	21.7%	18.8%	23.9%
partial (n=2)	0.0%	4.3%	0.0%	2.2%
yes (n=68)	70.0%	73.9%	81.3%	73.9%

Cramer's V = 0.1258

Target and actual sample sizes reported and non-responses bias discussed

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=42)	46.7%	39.1%	62.5%	45.7%
partial (n=40)	30.0%	54.3%	37.5%	43.5%
yes (n=10)	23.3%	6.5%	0.0%	10.9%

Cramer's V = 0.2433

Reliability: Conclusion

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
adequate (n=34)	33.3%	34.8%	50.0%	37.0%
not adequate (n=58)	66.7%	65.2%	50.0%	63.0%

Cramer's V = 0.1257

Cogency

Answers to all study questions, including sub-questions, included

	Low (n = 22)	Lower-Middle (n = 37)	Upper-Middle (n = 15)	Total (n = 74)
no (n=14)	18.2%	21.6%	13.3%	18.9%
yes (n=60)	81.8%	78.4%	86.7%	81.1%

Cramer's V = 0.0813
NA = Low (8), Lower-Middle (9), Upper-Middle (1)

Answers to all study questions included in the Executive Summary

	Low (n = 22)	Lower-Middle (n = 37)	Upper-Middle (n = 15)	Total (n = 74)
no (n=21)	27.3%	32.4%	20.0%	28.4%
yes (n=53)	72.7%	67.6%	80.0%	71.6%

Cramer's V = 0.1059
NA = Low (8), Lower-Middle (9), Upper-Middle (1)

Written in a language adequate to its stated audience

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=7)	10.0%	6.5%	6.3%	7.6%
yes (n=85)	90.0%	93.5%	93.8%	92.4%

Cramer's V = 0.0628

Recommendations are action-oriented, practical and specific

	Low (n = 29)	Lower-Middle (n = 41)	Upper-Middle (n = 16)	Total (n = 86)
no (n=6)	3.4%	12.2%	0.0%	7.0%
partial (n=33)	37.9%	31.7%	56.3%	38.4%
yes (n=47)	58.6%	56.1%	43.8%	54.7%

Cramer's V = 0.1787
NA = Low (1), Lower-Middle (5)

Connection between study questions, findings, conclusions and recommendations

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=5)	6.7%	2.2%	12.5%	5.4%
partial (n=43)	33.3%	52.2%	56.3%	46.7%
yes (n=44)	60.0%	45.7%	31.3%	47.8%

Cramer's V = 0.1821

Visuals are helpful for a non-technical audience to understand the findings

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
no (n=11)	16.7%	8.7%	12.5%	12.0%
partial (n=40)	30.0%	52.2%	43.8%	43.5%
yes (n=41)	53.3%	39.1%	43.8%	44.6%

Cramer's V = 0.1446

Cogency: Conclusion

	Low (n = 30)	Lower-Middle (n = 46)	Upper-Middle (n = 16)	Total (n = 92)
adequate (n=69)	66.7%	78.3%	81.3%	75.0%
not adequate (n=23)	33.3%	21.7%	18.8%	25.0%

Cramer's V = 0.1361

Results by Crisis and Conflict Environment

Conceptual Framing

Study questions included

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=18)	26.5%	15.5%	19.6%
yes (n=74)	73.5%	84.5%	80.4%

Cramer's V = 0.1333

Study hypotheses included

	Crisis and Conflict (n = 28)	Not Crisis and Conflict (n = 58)	Total (n = 86)
no (n=60)	67.9%	70.7%	69.8%
yes (n=26)	32.1%	29.3%	30.2%

Cramer's V = -0.0289
NA: Crisis and Conflict (6)

Study questions appropriate given the intervention's conceptual framework

	Crisis and Conflict (n = 22)	Not Crisis and Conflict (n = 49)	Total (n = 71)
no (n=2)	4.5%	2.0%	2.8%
partial (n=21)	18.2%	34.7%	29.6%
yes (n=48)	77.3%	63.3%	67.6%

Cramer's V = 0.1752
NA: Crisis and Conflict (12), Not in Crisis and Conflict (9)

Study acknowledges/draws upon existing country-specific research

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=25)	32.4%	24.1%	27.2%
partial (n=35)	26.5%	44.8%	38.0%
yes (n=32)	41.2%	31.0%	34.8%

Cramer's V = 0.1825

Local context provided allows non-experts to appreciate relevance of the study

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=18)	23.5%	17.2%	19.6%
partial (n=32)	32.4%	36.2%	34.8%
yes (n=42)	44.1%	46.6%	45.7%

Cramer's V = 0.0775

Conceptual Framing: Conclusion

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
adequate (n=62)	61.8%	70.7%	67.4%
not adequate (n=30)	38.2%	29.3%	32.6%

Cramer's V = -0.0919

Openness and Transparency

Open about limitations to implementing the study

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=11)	17.6%	8.6%	12.0%
partial (n=43)	55.9%	41.4%	46.7%
yes (n=38)	26.5%	50.0%	41.3%

Cramer's V = 0.2400

Open about limitations to implementing the intervention

	Crisis and Conflict (n = 28)	Not Crisis and Conflict (n = 57)	Total (n = 85)
no (n=17)	10.7%	24.6%	20.0%
partial (n=31)	46.4%	31.6%	36.5%
yes (n=37)	42.9%	43.9%	43.5%

Cramer's V = 0.1860

NA: Crisis and Conflict (6), Not in Crisis and Conflict (1)

Alternative interpretations of the findings included

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=72)	79.4%	77.6%	78.3%
yes (n=20)	20.6%	22.4%	21.7%

Cramer's V = 0.0214

Comprehensive analysis of the data relevant for study questions included

	Crisis and Conflict (n = 31)	Not Crisis and Conflict (n = 55)	Total (n = 86)
no (n=10)	16.1%	9.1%	11.6%
partial (n=36)	54.8%	34.5%	41.9%
yes (n=40)	29.0%	56.4%	46.5%

Cramer's V = 0.2637

NA: Crisis and Conflict (3), Not in Crisis and Conflict (3)

Open about potential biases due to the study team composition

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=62)	70.6%	65.5%	67.4%
partial (n=20)	23.5%	20.7%	21.7%
yes (n=10)	5.9%	13.8%	10.9%

Cramer's V = 0.1232

Openness and Transparency: Conclusion

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
adequate (n=45)	41.2%	53.4%	48.9%
not adequate (n=47)	58.8%	46.6%	51.1%

Cramer's V = -0.1185

Robustness of the Methodology

Methodology explained in detail

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=5)	5.9%	5.2%	5.4%
partial (n=41)	44.1%	44.8%	44.6%
yes (n=46)	50.0%	50.0%	50.0%

Cramer's V = 0.0156

Methodology appropriate for the study

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=6)	8.8%	5.2%	6.5%
partial (n=40)	47.1%	41.4%	43.5%
yes (n=46)	44.1%	53.4%	50.0%

Cramer's V = 0.1027

Counterfactual meets standards of rigor

	Crisis and Conflict (n = 10)	Not Crisis and Conflict (n = 17)	Total (n = 27)
no (n=6)	20.0%	23.5%	22.2%
yes (n=21)	80.0%	76.5%	77.8%

Cramer's V = -0.0410

NA: Crisis and Conflict (24), Not in Crisis and Conflict (41)

Data triangulation described as part of methodology

	Crisis and Conflict (n = 24)	Not Crisis and Conflict (n = 41)	Total (n = 65)
no (n=9)	8.3%	17.1%	13.8%
partial (n=23)	33.3%	36.6%	35.4%
yes (n=33)	58.3%	46.3%	50.8%

Cramer's V = 0.1419

NA: Crisis and Conflict (10), Not in Crisis and Conflict (17)

Addressed internal validity, either threats to inference or common biases

	Crisis and Conflict (n = 29)	Not Crisis and Conflict (n = 53)	Total (n = 82)
no (n=28)	34.5%	34.0%	34.1%
partial (n=38)	48.3%	45.3%	46.3%
yes (n=16)	17.2%	20.8%	19.5%

Cramer's V = 0.0437

NA: Crisis and Conflict (5), Not in Crisis and Conflict (5)

Described sampling approach and parameters used to compute sample size

	Crisis and Conflict (n = 14)	Not Crisis and Conflict (n = 26)	Total (n = 40)
no (n=9)	28.6%	19.2%	22.5%
partial (n=11)	21.4%	30.8%	27.5%
yes (n=20)	50.0%	50.0%	50.0%

Cramer's V = 0.1266
NA: Crisis and Conflict (20), Not in Crisis and Conflict (32)

Described sampling approach to collect qualitative data

	Crisis and Conflict (n = 18)	Not Crisis and Conflict (n = 31)	Total (n = 49)
no (n=11)	27.8%	19.4%	22.4%
partial (n=23)	50.0%	45.2%	46.9%
yes (n=15)	22.2%	35.5%	30.6%

Cramer's V = 0.1478
NA: Crisis and Conflict (16), Not in Crisis and Conflict (27)

Robustness of Methodology: Conclusion

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
adequate (n=60)	55.9%	70.7%	65.2%
not adequate (n=32)	44.1%	29.3%	34.8%

Cramer's V = -0.1501

Cultural Appropriateness

Included documentation from ethics review for approval/exemption status

	Crisis and Conflict (n = 32)	Not Crisis and Conflict (n = 57)	Total (n = 89)
no (n=79)	84.4%	91.2%	88.8%
yes (n=10)	15.6%	8.8%	11.2%

Cramer's V = -0.1041
NA: Crisis and Conflict (2), Not in Crisis and Conflict (1)

Study questions and methodology informed by local stakeholders

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=73)	70.6%	84.5%	79.3%
yes (n=19)	29.4%	15.5%	20.7%

Cramer's V = -0.1657

Data collection tools developed with participation of local stakeholders

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=56)	61.8%	60.3%	60.9%
partial (n=22)	23.5%	24.1%	23.9%
yes (n=14)	14.7%	15.5%	15.2%

Cramer's V = 0.0146

Findings/conclusions/recommendations validated with local stakeholders

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=81)	88.2%	87.9%	88.0%
yes (n=11)	11.8%	12.1%	12.0%

Cramer's V = 0.0045

Findings disaggregated by locally relevant stratifiers

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=15)	14.7%	17.2%	16.3%
partial (n=46)	52.9%	48.3%	50.0%
yes (n=31)	32.4%	34.5%	33.7%

Cramer's V = 0.0474

Cultural Appropriateness: Conclusion

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
adequate (n=27)	35.3%	25.9%	29.3%
not adequate (n=65)	64.7%	74.1%	70.7%

Cramer's V = 0.1000

Validity

Addressed construct validity of the assessment tools

	Crisis and Conflict (n = 32)	Not Crisis and Conflict (n = 54)	Total (n = 86)
partial (n=6)	14.3%	15.4%	15.0%
yes (n=34)	85.7%	84.6%	85.0%
partial (n=6)	14.3%	15.4%	15.0%

Cramer's V = -0.0147

NA: Crisis and Conflict (2), Not in Crisis and Conflict (4)

Addressed the external validity of findings from the sample to population

	Crisis and Conflict (n = 14)	Not Crisis and Conflict (n = 26)	Total (n = 40)
no (n=12)	42.9%	23.1%	30.0%
partial (n=8)	7.1%	26.9%	20.0%
yes (n=20)	50.0%	50.0%	50.0%

Cramer's V = 0.2724

NA: Crisis and Conflict (20), Not in Crisis and Conflict (32)

Addressed ecological validity of findings

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=80)	85.3%	87.9%	87.0%
yes (n=12)	14.7%	12.1%	13.0%

Cramer's V = -0.0378

Addressed the external validity of findings to other contexts

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=53)	52.9%	60.3%	57.6%
partial (n=24)	29.4%	24.1%	26.1%
yes (n=15)	17.6%	15.5%	16.3%

Cramer's V = 0.0731

Data collection tools piloted with representatives of target populations

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=56)	67.6%	56.9%	60.9%
partial (n=23)	20.6%	27.6%	25.0%
yes (n=13)	11.8%	15.5%	14.1%

Cramer's V = 0.1063

Confidence intervals reported around point estimates

	Crisis and Conflict (n = 11)	Not Crisis and Conflict (n = 23)	Total (n = 34)
no (n=18)	36.4%	60.9%	52.9%
yes (n=16)	63.6%	39.1%	47.1%

Cramer's V = -0.2297
NA: Crisis and Conflict (23), Not in Crisis and Conflict (35)

Relationships tested for statistical significance and *p*-value reported

	Crisis and Conflict (n = 11)	Not Crisis and Conflict (n = 23)	Total (n = 34)
yes (n=34)	100%	100%	100%

Cramer's V = NA
NA: Crisis and Conflict (23), Not in Crisis and Conflict (35)

Treatment effects presented in terms of effect sizes

	Crisis and Conflict (n = 10)	Not Crisis and Conflict (n = 23)	Total (n = 33)
no (n=14)	40.0%	43.5%	42.4%
yes (n=19)	60.0%	56.5%	57.6%

Cramer's V = -0.0323
NA: Crisis and Conflict (24), Not in Crisis and Conflict (35)

Validity: Conclusion

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
adequate (n=38)	32.4%	46.6%	41.3%
not adequate (n=54)	67.6%	53.4%	58.7%

Cramer's V = -0.1392

Reliability

Steps taken to ensure that data were reliably collected

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=41)	38.2%	48.3%	44.6%
partial (n=35)	52.9%	29.3%	38.0%
yes (n=16)	8.8%	22.4%	17.4%

Cramer's V = 0.2534

Inter-rater reliability statistics of assessors' fieldwork provided

	Crisis and Conflict (n = 14)	Not Crisis and Conflict (n = 26)	Total (n = 40)
no (n=36)	92.9%	88.5%	90.0%
yes (n=4)	7.1%	11.5%	10.0%

Cramer's V = 0.0699
NA: Crisis and Conflict (20), Not in Crisis and Conflict (32)

Data collection tools included in annex

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=22)	35.3%	17.2%	23.9%
partial (n=2)	2.9%	1.7%	2.2%
yes (n=68)	61.8%	81.0%	73.9%

Cramer's V = 0.2122

Target and actual sample sizes reported and non-responses bias discussed

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=42)	52.9%	41.4%	45.7%
partial (n=40)	29.4%	51.7%	43.5%
yes (n=10)	17.6%	6.9%	10.9%

Cramer's V = 0.2414

Reliability: Conclusion

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
adequate (n=34)	29.4%	41.4%	37.0%
not adequate (n=58)	70.6%	58.6%	63.0%

Cramer's V = -0.1197

Cogency

Answers to all study questions, including sub-questions, included

	Crisis and Conflict (n = 25)	Not Crisis and Conflict (n = 49)	Total (n = 74)
no (n=14)	24.0%	16.3%	18.9%
yes (n=60)	76.0%	83.7%	81.1%

Cramer's V = 0.0927
NA: Crisis and Conflict (9), Not in Crisis and Conflict (9)

Answers to all study questions included in the Executive Summary

	Crisis and Conflict (n = 25)	Not Crisis and Conflict (n = 49)	Total (n = 74)
no (n=21)	24.0%	30.6%	28.4%
yes (n=53)	76.0%	69.4%	71.6%

Cramer's V = -0.0694
NA: Crisis and Conflict (9), Not in Crisis and Conflict (9)

Written in a language adequate to its stated audience

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=7)	8.8%	6.9%	7.6%
yes (n=85)	91.2%	93.1%	92.4%

Cramer's V = 0.0351

Recommendations are action-oriented, practical and specific

	Crisis and Conflict (n = 32)	Not Crisis and Conflict (n = 54)	Total (n = 86)
no (n=6)	3.1%	9.3%	7.0%
partial (n=33)	53.1%	29.6%	38.4%
yes (n=47)	43.8%	61.1%	54.7%

Cramer's V = 0.2431

NA: Crisis and Conflict (2), Not in Crisis and Conflict (4)

Connection between study questions, findings, conclusions and recommendations

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=5)	0.0%	8.6%	5.4%
partial (n=43)	61.8%	37.9%	46.7%
yes (n=44)	38.2%	53.4%	47.8%

Cramer's V = 0.2673

Visuals are helpful for a non-technical audience to understand the findings

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
no (n=11)	5.9%	15.5%	12.0%
partial (n=40)	32.4%	50.0%	43.5%
yes (n=41)	61.8%	34.5%	44.6%

Cramer's V = 0.2715

Cogency: Conclusion

	Crisis and Conflict (n = 34)	Not Crisis and Conflict (n = 58)	Total (n = 92)
adequate (n=69)	70.6%	77.6%	75.0%
not adequate (n=23)	29.4%	22.4%	25.0%

Cramer's V = -0.0780

Results by Primary Education Strategy Goal

Conceptual Framing

Study questions included

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=18)	26.7%	19.6%	6.3%	19.6%
yes (n=74)	73.3%	80.4%	93.8%	80.4%

Cramer's V = 0.1849

Study hypotheses included

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 17)	Total (n = 86)
no (n=60)	79.3%	70.0%	52.9%	69.8%
yes (n=26)	20.7%	30.0%	47.1%	30.2%

Cramer's V = 0.2027
NA = Goal 3 (6)

Study questions appropriate given the intervention's conceptual framework

	Goal 1 (n = 22)	Goal 2 (n = 37)	Goal 3 (n = 12)	Total (n = 71)
no (n=2)	4.5%	2.7%	0.0%	2.8%
partial (n=21)	36.4%	32.4%	8.3%	29.6%
yes (n=48)	59.1%	64.9%	91.7%	67.6%

Cramer's V = 0.1710
NA = Goal 1 (7), Goal 2 (3), Goal 3 (11)

Study acknowledges/draws upon existing country-specific research

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=25)	24.1%	37.5%	13.0%	27.2%
partial (n=35)	48.3%	32.5%	34.8%	38.0%
yes (n=32)	27.6%	30.0%	52.2%	34.8%

Cramer's V = 0.1983

Local context provided allows non-experts to appreciate relevance of the study

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=18)	24.1%	25.0%	4.3%	19.6%
partial (n=32)	37.9%	42.5%	17.4%	34.8%
yes (n=42)	37.9%	32.5%	78.3%	45.7%

Cramer's V = 0.2724

Conceptual Framing: Conclusion

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
adequate (n=62)	65.5%	70.0%	65.2%	67.4%
not adequate (n=30)	34.5%	30.0%	34.8%	32.6%

Cramer's V = 0.0489

Openness and Transparency

Open about limitations to implementing the study

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=11)	6.9%	5.0%	30.4%	12.0%
partial (n=43)	48.3%	47.5%	43.5%	46.7%
yes (n=38)	44.8%	47.5%	26.1%	41.3%

Cramer's V = 0.2403

Open about limitations to implementing the intervention

	Goal 1 (n = 29)	Goal 2 (n = 39)	Goal 3 (n = 17)	Total (n = 85)
no (n=17)	17.2%	25.6%	11.8%	20.0%
partial (n=31)	37.9%	38.5%	29.4%	36.5%
yes (n=37)	44.8%	35.9%	58.8%	43.5%

Cramer's V = 0.1338
NA = Goal 2 (1), Goal 3 (6)

Alternative interpretations of the findings included

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=72)	75.9%	90.0%	60.9%	78.3%
yes (n=20)	24.1%	10.0%	39.1%	21.7%

Cramer's V = 0.2841

Comprehensive analysis of the data relevant for study questions included

	Goal 1 (n = 27)	Goal 2 (n = 39)	Goal 3 (n = 20)	Total (n = 86)
no (n=10)	22.2%	5.1%	10.0%	11.6%
partial (n=36)	25.9%	53.8%	40.0%	41.9%
yes (n=40)	51.9%	41.0%	50.0%	46.5%

Cramer's V = 0.2093
NA = Goal 1 (2), Goal 2 (1), Goal 3 (3)

Open about potential biases due to the study team composition

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=62)	62.1%	67.5%	73.9%	67.4%
partial (n=20)	24.1%	22.5%	17.4%	21.7%
yes (n=10)	13.8%	10.0%	8.7%	10.9%

Cramer's V = 0.0703

Openness and Transparency: Conclusion

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
adequate (n=45)	55.2%	47.5%	43.5%	48.9%
not adequate (n=47)	44.8%	52.5%	56.5%	51.1%

Cramer's V = 0.0908

Robustness of the Methodology

Methodology explained in detail

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=5)	3.4%	7.5%	4.3%	5.4%
partial (n=41)	55.2%	45.0%	30.4%	44.6%
yes (n=46)	41.4%	47.5%	65.2%	50.0%

Cramer's V = 0.1453

Methodology appropriate for the study

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=6)	6.9%	10.0%	0.0%	6.5%
partial (n=40)	41.4%	45.0%	43.5%	43.5%
yes (n=46)	51.7%	45.0%	56.5%	50.0%

Cramer's V = 0.1214

Counterfactual meets standards of rigor

	Goal 1 (n = 12)	Goal 2 (n = 4)	Goal 3 (n = 11)	Total (n = 27)
no (n=6)	41.7%	0.0%	9.1%	22.2%
yes (n=21)	58.3%	100%	90.9%	77.8%

Cramer's V = 0.4245
NA = Goal 1 (17), Goal 2 (36), Goal 3 (12)

Data triangulation described as part of methodology

	Goal 1 (n = 17)	Goal 2 (n = 36)	Goal 3 (n = 12)	Total (n = 65)
no (n=9)	23.5%	11.1%	8.3%	13.8%
partial (n=23)	47.1%	33.3%	25.0%	35.4%
yes (n=33)	29.4%	55.6%	66.7%	50.8%

Cramer's V = 0.1953
NA = Goal 1 (12), Goal 2 (4), Goal 3 (11)

Addressed internal validity, either threats to inference or common biases

	Goal 1 (n = 26)	Goal 2 (n = 38)	Goal 3 (n = 18)	Total (n = 82)
no (n=28)	30.8%	42.1%	22.2%	34.1%
partial (n=38)	50.0%	39.5%	55.6%	46.3%
yes (n=16)	19.2%	18.4%	22.2%	19.5%

Cramer's V = 0.1216
NA = Goal 1 (3), Goal 2 (2), Goal 3 (5)

Described sampling approach and parameters used to compute sample size

	Goal 1 (n = 14)	Goal 2 (n = 15)	Goal 3 (n = 11)	Total (n = 40)
no (n=9)	21.4%	33.3%	9.1%	22.5%
partial (n=11)	35.7%	20.0%	27.3%	27.5%
yes (n=20)	42.9%	46.7%	63.6%	50.0%

Cramer's V = 0.1905
NA = Goal 1 (15), Goal 2 (25), Goal 3 (12)

Described sampling approach to collect qualitative data

	Goal 1 (n = 14)	Goal 2 (n = 24)	Goal 3 (n = 11)	Total (n = 49)
no (n=11)	21.4%	20.8%	27.3%	22.4%
partial (n=23)	57.1%	45.8%	36.4%	46.9%
yes (n=15)	21.4%	33.3%	36.4%	30.6%

Cramer's V = 0.1147
NA = Goal 1 (15), Goal 2 (16), Goal 3 (12)

Robustness of Methodology: Conclusion

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
adequate (n=60)	72.4%	62.5%	60.9%	65.2%
not adequate (n=32)	27.6%	37.5%	39.1%	34.8%

Cramer's V = 0.1034

Cultural Appropriateness

Included documentation from ethics review for approval/exemption status

	Goal 1 (n = 29)	Goal 2 (n = 38)	Goal 3 (n = 22)	Total (n = 89)
no (n=79)	89.7%	97.4%	72.7%	88.8%
yes (n=10)	10.3%	2.6%	27.3%	11.2%

Cramer's V = 0.3093
NA = Goal 2 (2), Goal 3 (1)

Study questions and methodology informed by local stakeholders

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=73)	82.8%	85.0%	65.2%	79.3%
yes (n=19)	17.2%	15.0%	34.8%	20.7%

Cramer's V = 0.2029

Data collection tools developed with participation of local stakeholders

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=56)	55.2%	62.5%	65.2%	60.9%
partial (n=22)	31.0%	20.0%	21.7%	23.9%
yes (n=14)	13.8%	17.5%	13.0%	15.2%

Cramer's V = 0.0875

Findings/conclusions/recommendations validated with local stakeholders

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=81)	86.2%	87.5%	91.3%	88.0%
yes (n=11)	13.8%	12.5%	8.7%	12.0%

Cramer's V = 0.0605

Findings disaggregated by locally relevant stratifiers

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=15)	27.6%	17.5%	0.0%	16.3%
partial (n=46)	44.8%	55.0%	47.8%	50.0%
yes (n=31)	27.6%	27.5%	52.2%	33.7%

Cramer's V = 0.2276

Cultural Appropriateness: Conclusion

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
adequate (n=27)	37.9%	17.5%	39.1%	29.3%
not adequate (n=65)	62.1%	82.5%	60.9%	70.7%

Cramer's V = 0.2284

Validity

Addressed construct validity of the assessment tools

	Goal 1 (n = 14)	Goal 2 (n = 15)	Goal 3 (n = 11)	Total (n = 40)
partial (n=6)	14.3%	26.7%	0.0%	15.0%
yes (n=34)	85.7%	73.3%	100%	85.0%
partial (n=6)	14.3%	26.7%	0.0%	15.0%

Cramer's V = 0.2978
NA = Goal 1 (15), Goal 2 (25), Goal 3 (12)

Addressed the external validity of findings from the sample to population

	Goal 1 (n = 14)	Goal 2 (n = 15)	Goal 3 (n = 11)	Total (n = 40)
no (n=12)	7.1%	60.0%	18.2%	30.0%
partial (n=8)	42.9%	13.3%	0.0%	20.0%
yes (n=20)	50.0%	26.7%	81.8%	50.0%

Cramer's V = 0.4677
NA = Goal 1 (15), Goal 2 (25), Goal 3 (12)

Addressed ecological validity of findings

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=80)	89.7%	82.5%	91.3%	87.0%
yes (n=12)	10.3%	17.5%	8.7%	13.0%

Cramer's V = 0.1175

Addressed the external validity of findings to other contexts

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=53)	62.1%	67.5%	34.8%	57.6%
partial (n=24)	24.1%	25.0%	30.4%	26.1%
yes (n=15)	13.8%	7.5%	34.8%	16.3%

Cramer's V = 0.2322

Data collection tools piloted with representatives of target populations

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=56)	48.3%	67.5%	65.2%	60.9%
partial (n=23)	34.5%	17.5%	26.1%	25.0%
yes (n=13)	17.2%	15.0%	8.7%	14.1%

Cramer's V = 0.1432

Confidence intervals reported around point estimates

	Goal 1 (n = 14)	Goal 2 (n = 9)	Goal 3 (n = 11)	Total (n = 34)
no (n=18)	71.4%	33.3%	45.5%	52.9%
yes (n=16)	28.6%	66.7%	54.5%	47.1%

Cramer's V = 0.3234
NA = Goal 1 (15), Goal 2 (31), Goal 3 (12)

Relationships tested for statistical significance and *p*-value reported

	Goal 1 (n = 14)	Goal 2 (n = 9)	Goal 3 (n = 11)	Total (n = 34)
yes (n=34)	100%	100%	100%	100%

Cramer's V = NA
NA = Goal 1 (15), Goal 2 (31), Goal 3 (12)

Treatment effects presented in terms of effect sizes

	Goal 1 (n = 14)	Goal 2 (n = 8)	Goal 3 (n = 11)	Total (n = 33)
no (n=14)	28.6%	62.5%	45.5%	42.4%
yes (n=19)	71.4%	37.5%	54.5%	57.6%

Cramer's V = 0.2731
NA = Goal 1 (15), Goal 2 (32), Goal 3 (12)

Validity: Conclusion

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
adequate (n=38)	51.7%	27.5%	52.2%	41.3%
not adequate (n=54)	48.3%	72.5%	47.8%	58.7%

Cramer's V = 0.2459

Reliability

Steps taken to ensure that data were reliably collected

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=41)	51.7%	40.0%	43.5%	44.6%
partial (n=35)	24.1%	45.0%	43.5%	38.0%
yes (n=16)	24.1%	15.0%	13.0%	17.4%

Cramer's V = 0.1442

Inter-rater reliability statistics of assessors' fieldwork provided

	Goal 1 (n = 14)	Goal 2 (n = 15)	Goal 3 (n = 11)	Total (n = 40)
no (n=36)	78.6%	93.3%	100%	90.0%
yes (n=4)	21.4%	6.7%	0.0%	10.0%

Cramer's V = 0.2932
NA = Goal 1 (15), Goal 2 (25), Goal 3 (12)

Data collection tools included in annex

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=22)	24.1%	15.0%	39.1%	23.9%
partial (n=2)	3.4%	2.5%	0.0%	2.2%
yes (n=68)	72.4%	82.5%	60.9%	73.9%

Cramer's V = 0.1686

Target and actual sample sizes reported and non-responses bias discussed

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=42)	44.8%	52.5%	34.8%	45.7%
partial (n=40)	44.8%	42.5%	43.5%	43.5%
yes (n=10)	10.3%	5.0%	21.7%	10.9%

Cramer's V = 0.1616

Reliability: Conclusion

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
adequate (n=34)	37.9%	42.5%	26.1%	37.0%
not adequate (n=58)	62.1%	57.5%	73.9%	63.0%

Cramer's V = 0.1362

Cogency

Answers to all study questions, including sub-questions, included

	Goal 1 (n = 22)	Goal 2 (n = 37)	Goal 3 (n = 15)	Total (n = 74)
no (n=14)	27.3%	18.9%	6.7%	18.9%
yes (n=60)	72.7%	81.1%	93.3%	81.1%

Cramer's V = 0.1827
NA = Goal 1 (7), Goal 2 (3), Goal 3 (8)

Answers to all study questions included in the Executive Summary

	Goal 1 (n = 22)	Goal 2 (n = 37)	Goal 3 (n = 15)	Total (n = 74)
no (n=21)	36.4%	27.0%	20.0%	28.4%
yes (n=53)	63.6%	73.0%	80.0%	71.6%

Cramer's V = 0.1285
NA = Goal 1 (7), Goal 2 (3), Goal 3 (8)

Written in a language adequate to its stated audience

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=7)	6.9%	7.5%	8.7%	7.6%
yes (n=85)	93.1%	92.5%	91.3%	92.4%

Cramer's V = 0.0256

Recommendations are action-oriented, practical and specific

	Goal 1 (n = 28)	Goal 2 (n = 39)	Goal 3 (n = 19)	Total (n = 86)
no (n=6)	10.7%	5.1%	5.3%	7.0%
partial (n=33)	25.0%	46.2%	42.1%	38.4%
yes (n=47)	64.3%	48.7%	52.6%	54.7%

Cramer's V = 0.1439
NA = Goal 1 (1), Goal 2 (1), Goal 3 (4)

Connection between study questions, findings, conclusions and recommendations

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=5)	10.3%	5.0%	0.0%	5.4%
partial (n=43)	31.0%	50.0%	60.9%	46.7%
yes (n=44)	58.6%	45.0%	39.1%	47.8%

Cramer's V = 0.1849

Visuals are helpful for a non-technical audience to understand the findings

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
no (n=11)	24.1%	10.0%	0.0%	12.0%
partial (n=40)	37.9%	50.0%	39.1%	43.5%
yes (n=41)	37.9%	40.0%	60.9%	44.6%

Cramer's V = 0.2214

Cogency: Conclusion

	Goal 1 (n = 29)	Goal 2 (n = 40)	Goal 3 (n = 23)	Total (n = 92)
adequate (n=69)	72.4%	72.5%	82.6%	75.0%
not adequate (n=23)	27.6%	27.5%	17.4%	25.0%

Cramer's V = 0.1015

ANNEX 7: REFERENCES

Abt Associates. "EVIRATER: Rating the Strength of Evidence in Evaluations." Accessed September 1, 2017. http://abtassociates.com/Noteworthy/2015/EVIRATER-Rating-the-Strength-of-Evidence-in-Evalua.aspx

ADS 201 Additional Help: USAID Recommended Data Quality Assessment (DQA) Checklist. USAID, September 2016. https://www.usaid.gov/sites/default/files/documents/1870/201maa.pdf

Building Evidence in Education (BE^2) Steering Committee. *Assessing the Strength of Evidence in the Education Sector.* 2015. https://www.usaid.gov/sites/default/files/documents/1865/BE2_Guidance_Note_ASE.pdf

Creswell, John W., and Vicki L. Plano Clark. *Designing and Conducting Mixed Methods Research*. California: SAGE Publications, 2011.

Criteria to Ensure the Quality of the Evaluation Report: A Mandatory Reference for ADS Chapter 201. USAID, September 7, 2016. https://www.usaid.gov/sites/default/files/documents/1870/201maa.pdf

Freudenberger, Elizabeth. *Sectoral Synthesis of 2013–2014 Evaluation Findings: Bureau for Economic Growth, Education, & Environment.* UDSID, August 2015. https://www.usaid.gov/sites/default/files/documents/1865/E3_Sectoral_Synthesis_Report.pdf

General Accountability Office (GAO). *Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluations.* March 2017. http://www.gao.gov/assets/690/683157.pdf

Glennerster, Rachel, and Kudzai Takavarasha, *Running Randomized Evaluations: A Practical Guide*. New Jersey: Princeton University Press, 2013.

Guest, Greg, Emily Namey, and Kevin McKenna. "How many focus groups are enough? Building an evidence base for nonprobability sample sizes." Field Methods 29, no. 1 (2017): 3-22

Green, Andrew, and Sam Hargadine. *Sectoral Synthesis of FY2015 Evaluation Findings: Bureau for Economic Growth, Education, and Environment.* USAID, December 2016. http://pdf.usaid.gov/pdf_docs/PA00MP17.pdf

Hageboeck, Molly; Micah Frumkin, Jenna L. Heavenrich, Lala Kasimova, Melvin M. Mark, and Aníbal Pérez-Liñán. *Evaluation Utilization at USAID*. USAID, February 23, 2016. http://pdf.usaid.gov/pdf_docs/PA00KXVT.pdf

Hageboeck, Molly, Micah Frumkin, and Stephanie Monschein. *Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009 – 2012.* USAID, August 2013. http://pdf.usaid.gov/pdf_docs/pdacx771.pdf

Management Systems International (MSI). *Evaluation Quality Rater's Guide – Pilot Version: Education Evaluation Syntheses*. USAID, June 26, 2017. https://s3.amazonaws.com/edeval-documents/resources/Evaluation+Quality+Raters+Guide+PILOT+6-26-17.pdf

Pritchett, Lant. *The Evidence About What Works in Education: Graphs to Illustrate External Validity and Construct Validity*. RISE Insights, June 2017. Retrieved from http://www.riseprogramme.org/content/evidence-about-what-works-education-graphs-illustrate-external-validity-and-construct

Qualitative Research Checklist. Critical Appraisal Skills Programme (CASP), March 13, 2017.

RTI International. *Early Grade Reading Assessment (EGRA) Toolkit: Second Edition*. USAID, March 2016. https://globalreadingnetwork.net/resources/early-grade-reading-assessment-egra-toolkit-second-edition

United States Department of Labor. "CLEAR: Clearinghouse for Labor Evaluation and Research." Accessed September 1, 2017. https://clear.dol.gov/
USAID Evaluation Report Requirements: A Mandatory Reference for ADS Chapter 201. USAID, September 7, 2016. https://www.usaid.gov/sites/default/files/documents/1870/201maa.pdf

USAID Evaluation Policy. USAID, October 2016.
https://www.usaid.gov/sites/default/files/documents/1870/USAIDEvaluationPolicy.pdf

USAID Scientific Research Policy. USAID, December 2014.
https://www.usaid.gov/sites/default/files/documents/1870/201maa.pdf

What Works Clearinghouse. *Procedures and Standards Handbook, Version 3.0*. 2014.
https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

U.S. Agency for International Development
1300 Pennsylvania Avenue, NW
Washington, DC 20004